UNIVERSITY OF CALIFORNIA
RIVERSIDE

A Stochastic Restricted Maximum Likelihood Method for Genomic Selection and Genome-Wide Association Studies

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Chen Lin

March 2019

Dissertation Committee:

Dr. Shizhong Xu, Chairperson
Dr. Weixin Yao
Dr. Zhenyu Jia

The Dissertation of Chen Lin is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

My most sincere thanks to my advisor, Dr. Shizhong Xu, for guiding and encouraging me through my Ph.D experience. His creativity and enthusiasm for researches have been fundamentally influenced me, which will guide me profoundly. I also want to express my gratitude towards my co-advisor Dr. Weixin Yao and my committee member Dr. Zhenyu Jia for their support and valuable suggestions. I would like to thank all the faculty and staff in the department for their help. I am grateful to my lab members, Ruidong Li, Han Qu, Meiyue Wang, Shibo Wang, Fangjie Xie, Tiantian Zhu for their support and friendship. I am extremely thankful to my friends Hua Peng, Lijie Li for their companionship. I would also like to thank all my other friends who have supported me along the way. Lastly, I sincerely thank my parents for their unconditional love and encouragement.

To my beloved parents.

ABSTRACT OF THE DISSERTATION

A Stochastic Restricted Maximum Likelihood Method for Genomic Selection and Genome-Wide Association Studies

by

Chen Lin

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, March 2019
Dr. Shizhong Xu, Chairperson

Genomic selection is a marker-assisted methodology that dramatically decreases the cost of measuring phenotypes by using the whole-genome information to predict and select desirable individuals. In plant breeding, it plays an important role to speed up the breeding cycles. Modern techniques make obtaining marker information from the entire genome feasible. However, it results in high dimensionality of predictors when we implement a mathematical model to estimate the parameters and predict future crosses. Many statistical models including variable selection models can address this problem and have been applied in genomic selection. Variable selection models can also be applied in GWAS which is a powerful tool to discover the association between genetic variation and variation in quantitative traits.

A novel statistical approach based on BLUP was proposed to be implemented in both genomic selection and GWAS. The general idea of the proposed approach is using an algorithm to divide markers into the small effect group and the large effect group. Markers within the large effect group can be potentially significant markers associated with the

analyzed phenotypic trait. In Chapter 3, we used simulated data and two real-world data sets to demonstrate the distinctions among six statistical methods for genomic selection. In addition, the proposed model was applied in GWAS based on another simulated data, and the proposed model is superior to the other two variable selection models.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Genomic Selection

A long time ago, people had to spend a lot of months or even years to obtain desirable animals or plants. Since 1990s, marker-assisted selection (MAS) has been used to indirectly select desirable candidates with the genetic markers significantly associated with a particular trait. However, this methodology is time-consuming attributed to the long breeding cycles and genetic information loss. In contrast, genomic selection provides a way to overcome these limitations. It is well-known as a marker-assisted methodology that drastically reduces the cost of measuring quantitative traits by using the whole-genome information to predict and directly select desirable individuals. Here selection stands for selecting future individuals based on the predicted phenotypes rather than selecting genetic markers associated with a quantitative trait. It is particularly useful to accelerate the breeding process. Moreover, it facilitates to predict phenotypic values for a breeding population (they are only genotyped) according to the model fitted in a training population

(they are both genotyped and phenotyped) (Desta and Ortiz, 2014). Genomic selection has been applied in many areas, such as human beings, as well as animal and plant species. In next subsection, we introduce the applications of genomic selection in these areas.

1.1.1 Applications of Genomic Selection

Genomic selection has been widely used since 2001. Meuwissen et al. (2001) attempted to use the whole genomic information to build the linear regression model and predict breeding values in a simulated data set based on animal genotypes under BLUP. They also proposed new methods (BayesA and BayesB) and compared the results drawn from new methods with BLUP. The limitations of genomic selection are the requirement of a large number of markers and the cost of marker discovery. Fortunately, these problems were addressed a few years later. Hayes et al. (2009) investigated and summarized the accuracy of genomic breeding values (GEBV) in Holstein-Friesian dairy cattle from Australia, New Zealand, and the United States (Harris et al., 2009; VanRaden et al., 2009). A total of 4,500 bulls and a total of 44,146 markers were involved in New Zealand dairy cattle analysis. A total of 730 bulls and a total of 38,259 markers were involved in Australia dairy cattle analysis. The application of dairy cattle in the United States consisted of a total of 3,576 bulls and a total of 38,416 markers. New Zealand and America dairy cattle populations achieved the similar accuracies, and Australia dairy cattle data had the lowest accuracy. Erbe et al. (2012) combined Holstein and Jersey reference populations to improve prediction accuracy in Jersey cattle. Besides dairy cattle, genomic selection was applied in other animal species as well. For instants, Legarra et al. (2008) examined three strategies of selection for four complex traits in a mouse population including 1,884 mice

and 10,946 markers. Lee et al. (2008) proposed a Bayesian method and implemented in a heterogeneous stock mouse population to predict unobserved phenotypic values.

Genomic selection has been applied in animal species for a long time. However, it is in its infancy in plant breeding and humans. Bernardo and Yu (2007) compared results drawn from genomic selection with results drawn from MAS under BLUP in a maize population and concluded genomic selection outperformed MAS. Piepho (2009) investigated ridge regression and some other methods in maize. Heffner et al. (2009) summarized technologies used in genomic selection for crop improvement. Xu et al. (2014) exploited hybrid prediction in genomic selection. A total of 278 hybrids was used to predict 21,945 potential hybrids to select top crosses. Yang et al. (2010) uncovered the missing heritability for human height using whole-genome information.

1.2 Genome-wide Association Studies (GWAS)

In genetics, GWAS (Risch and Merikangas, 1996) is another crucial field. It facilitates to discover genetic variation associated with a complex or quantitative trait (including diseases) in a genome-wide panel of markers. A linear mixed model is widely used to identify the statistical associations between single-nucleotide polymorphisms (SNPs) and phenotypic traits. It consists of covariate effects, a single marker fixed effect, polygenic effects controlled by all other genes, and it is well-known as single-SNP tests, i.e. testing one SNP at a time.

GWAS provide a powerful tool to define phenotypic traits across individuals and disclose the causal relationship between genetic variants and phenotypic differences. Discovering the underlying genetic architecture provides insights into the understanding of

complex traits and disease susceptibility. It has been successfully applied in many areas, such as human diseases and economic traits in crops.

1.2.1 Applications of GWAS

Many human diseases have been investigated, and the associated SNPs have been uncovered. For coronary heart disease (CHD), Consortium et al. (2011) identified five new markers associated with CHD risk in Europeans and South Asians; Domarkienė et al. (2013) identified two important loci associated with CHD risk in Lithuanian Families. For type 2 diabetes mellitus (T2D), Sim et al. (2011) exploited significant loci associated with T2D in multi-ethnic cohorts including Chinese, Malays, and Asian Indians; Ghassibe-Sabbagh et al. (2014) identified two markers associated with T2D in the Lebanese population to verify the key role of these two loci in Southwest Asian populations. Li et al. (2015) analyzed 10 pediatric-age-of-onset autoimmune diseases (pAIDs) and revealed 27 significant markers related to one or more pAIDs. Michailidou et al. (2017) detected 65 new loci associated with breast cancer in a population including European ancestry and East Asian ancestry.

In crops, Huang et al. (2010) performed GWAS to investigate ~3.6 million loci across 517 diverse rice landraces for 14 agronomic traits including tiller number, grain weight, grain width, etc. They also proposed a novel data-imputation method to construct a high-density haplotype map. A total of 80 significant associations were detected for 14 agronomic traits. In 2011, 44,100 SNPs obtained from 413 diverse accessions of *Oryza sativa* were used to discover the relationships with 34 quantitative traits (Zhao et al., 2011). Dozens of common variants were identified to have influences on numerous traits. Jia et al. (2013) conducted GWAS of 47 agronomic traits based on ~2.58 million SNPs by sequencing 916 diverse foxtail

millet varieties. A total of 512 association signals had relationships with 47 agronomic traits. Chen et al. (2014) carried out GWAS to find the associations between genetic variants and metabolic traits by using˜6.4 million SNPs based on 529 diverse *Oryza sativa* accessions.

1.2.2 Single-SNP Tests

One of the most commonly used models in GWAS is linear mixed model (LMM) (Yu et al., 2006). The model for marker i can be written as

$$y = X\beta + z_i\gamma_i + \xi + e,$$

$$\xi \sim \text{MVN}_n(0, ZZ^T\sigma_\xi^2),$$

$$e \sim \text{MVN}_n(0, I\sigma_e^2)$$

where y is an $n \times 1$ vector of phenotypic values, X is a known $n \times q$ design matrix associated with fixed effects, β is a $q \times 1$ vector of unknown fixed effects, Z is a known $n \times m$ genetic matrix, z_i is the i^{th} column of Z, γ_i is an unknown fixed effect associated with i^{th} marker, ξ is the polygenic effect, e is an $n \times 1$ vector of random errors and $i = 1, \ldots, m$. MVN denotes a multivariate normal distribution. Every marker in the data set is required to be scanned to estimate its fixed marker effect γ_i. Unknown parameters can be estimated using REML method (introduced in Chapter 2).

The purpose of GWAS is to reveal markers significantly associated with a complex trait. Based on LMM, a hypothesis testing can be conducted to test the significance of the i^{th} marker, i.e. $\text{H}_0 : \gamma_i = 0$. The Wald test statistic for it is

$$W_i = \frac{\hat{\gamma}_i^2}{\text{Var}(\hat{\gamma}_i)}.$$

Under the null hypothesis, $\mathrm{H}_0 : \gamma_i = 0$, the distribution of W_i is $W_i \sim \chi^2_1$. Then the p-value of the marker i is

$$p_i = 1 - \Pr(\chi^2_1 \leq W_i).$$

For a single hypothesis test, p-value is usually compared with the level of significance α, also known as the type I error. If p-value is less than or equal to α, then we reject the null hypothesis; otherwise, the null hypothesis is failed to be rejected. However, for multiple comparisons (testing multiple hypotheses simultaneously), if we keep using the same criterion, the overall type I error, family-wise error rate (FWER), will vastly increase. FWER can be defined by

$$\mathrm{FWER} = \Pr(\mathrm{V} \geq 1),$$

where V is the number of false positives, i.e. how many true null hypotheses are rejected. If we assume hypotheses are independent of each other, then FWER can be calculated by $\mathrm{FWER} = 1 - (1 - \alpha)^m$. Suppose $m = 100$ and $\alpha = 0.05$, then FWER=0.99. It implies any m exceeding 100 leads to 100% probability that at least one null hypothesis will be rejected incorrectly. In order to address it, it is necessary to correct the level of significance in multiple comparisons.

One of the simplest and most popular correction methods is Bonferroni correction. It simply adjusts the level of significance by α/m, where m is the number of hypotheses. According to Boole's inequality, FWER has the property that it will be controlled within α if $p_i \leq \frac{\alpha}{m}$. Bonferroni correction assumes hypotheses are independent, but GWAS violate this assumption because of correlations existing among markers. Due to it, Bonferroni correction is conservative and leads to statistical power reduction.

In contrast to single-SNP tests, polygenic modeling with variable selection feature is an alternative way to identify important markers, such as LASSO, BayesB, etc. It estimates all marker coefficients simultaneously. Therefore, multiple comparisons are not involved in it. Here is an example of using LASSO in GWAS: Wu et al. (2009) implemented LASSO penalized logistic regression to identify important markers associated with coeliac diseases.

Details about polygenic modeling are introduced in Chapter 2.

1.3 The Stochastic EM Algorithm

In statistics, a lot of problems involves missing data or incomplete data. One of the most popular algorithms to solve these problems is the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The main idea of the EM algorithm is replacing maximization of the log likelihood function of the observed data with maximization of the conditional expectation of unobserved data.

Let us define the complete data as $x = (y, z)$ and it can be sampled from $f(x|\theta)$, where y denotes the observed data, z denotes the unobserved data or the latent variables, and θ denotes the unknown parameters. It's difficult to compute the likelihood function of the observed data, $f(y|\theta) = \int_z f(y, z|\theta)dz$, in many situations. Then the conditional expectation of unobserved data z given the observed data y and the estimated parameters $\theta^{(t)}$ is used and it is written as

$$Q(\theta|\theta^{(t)}) = \mathrm{E}_{z|y,\theta^{(t)}}(L(\theta; y, z)),$$

where $L(\theta; y, z)$ is the log likelihood function of the complete data.

The EM algorithm generates estimates as follows:

Step 1 (E step): Compute the conditional expectation of z, $Q(\theta|\theta^{(t)})$.

Step 2 (M step): Estimate unknown parameters $\theta^{(t+1)}$ by maximizing $Q(\theta|\theta^{(t)})$.

However, the EM algorithm is difficult to implement in some cases. Stochastic EM algorithms provide powerful tools to deal with the cases that can not be handled by the EM algorithm, including the SEM algorithm (Celeux, 1985), MCEM (Wei and Tanner, 1990), etc. In order to converge the estimates quickly, a stochastic EM algorithm is drawing the unobserved samples from the conditional density, $f(z|y,\theta)$. After that, the likelihood function can be directly maximized. This idea is implemented in our proposed method introduced in Chapter 3.

Chapter 2

Statistical Models Overview

With the remarkable advances in computing technology, millions of single-nucleotide polymorphisms (SNPs), a common type of genetic variation, can be obtained using modern genome sequencing technologies. Namely, high-density markers are involved in recent researches. In quantitative genetics, we treat SNPs or markers as our predictors and variation in quantitative traits of individuals as our response variable. The goals of statistical models are to find the relationship between our predictors and the response variable, and to estimate all coefficients simultaneously.

High-density marker panels provide more genetic information to improve the predictive accuracy. On the other hand, it also leads to the number of predictors substantially exceeding the number of individuals in a statistical model, and multicollinearity of predictors, i.e. high intercorrelations existing among several independent predictors. In ordinary least squares (OLS) regression, high dimensionality of markers results in a singular matrix which implies the solution is not unique, and multicollinearity causes the reduction of the statistical power. Fortunately, many other polygenic methodologies can deal with these

issues in statistics. In general, there are two types of these methodologies: linear regression models and nonlinear regression models. Specifically, we focus on linear models because of interpretability. Linear models dealing with high dimensionality issue mainly include regularized linear regression models, the linear mixed model, and Bayesian approaches.

2.1 Regularized Linear Regression Models

To solve high dimensionality problem of predictors, regularized linear regression models add a penalty term into the sum of squared residuals, i.e. penalized residual sum of squares, and then minimize it to estimate the unknown parameters. The general model can be written as follows

$$y = \mu + X\beta + e,$$

$$e \sim \text{MVN}_n(0, I\sigma_e^2),$$

where y is an $n \times 1$ vector of a response variable measured on n individuals, μ is an $n \times 1$ vector of intercept terms, X is an $n \times m$ design matrix ($m > n$), β is an $m \times 1$ vector of unknown parameters, e is an $n \times 1$ vector of random errors, and MVN_n represents the n-dimensional multivariate normal distribution.

If y and X are both centered, denoted by $\tilde{y}$ and $\tilde{X}$,then the unknown parameter β can be estimated with the following equation

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (\tilde{y}_i - \tilde{x}_i^T \beta)^2 + \lambda f(\beta),$$

where $\tilde{x}_i^T$ is the i^{th} row of $\tilde{X}$, $\lambda \geq 0$ is a shrinkage factor which controls the size of coefficients. The larger value of λ, the smaller value of β. $f(\cdot)$ represents the penalized function of β. Different functions lead to different regularized regression models. For instance, it derives

to ridge regression if the penalized function is l_2 penalty term, i.e. $f(\beta) = \|\beta\|_2^2 = \sum_{i=1}^{m} \beta_i^2$ (Hoerl and Kennard, 1970a,b); a famous regularized regression model, the least absolute shrinkage selection operator (LASSO), is proposed if the penalized function is identical to l_1 penalty term, i.e. $f(\beta) = \|\beta\|_1 = \sum_{i=1}^{m} |\beta_i|$ (Tibshirani, 1996).

There are two distinctions between two models. First, ridge regression shrinks all coefficients with equal sizes, while LASSO regression combines coefficient shrinkage and variable selection. l_1-norm has an additional property of shrinking some coefficients towards zero than l_2-norm. Second, ridge regression has an analytical solution, but LASSO regression does not, which implies that ridge regression is more statistically and computationally efficient than LASSO. Researchers have to develop faster algorithms to estimate LASSO parameters. Efron et al. (2004) developed least angle regression (LARS) algorithm to reduce runtime complexity of the algorithm to $O(nm^2)$ which is the same as OLS. LARS algorithm played a crucial role to calculation LASSO estimation before 2010. Friedman et al. (2010) introduced a simpler and more flexible algorithm, known as pathwise coordinate descent, to reduce the computational cost to $O(2nm)$. And this algorithm has still been widely used until now.

2.2 Linear Mixed Models

Another approach dealing with high dimensionality of predictors is linear mixed models, which is well-known as BLUP method. A linear mixed model can be expressed as

$$y = \mu + X\beta + Z\gamma + e,$$

$$\gamma \sim \mathrm{MVN}_m(0, G),$$

$$e \sim \text{MVN}_n(0, R),$$

where y is an $n \times 1$ observation vector, μ is an $n \times 1$ vector of intercept terms, X is a known $n \times q$ matrix associated with fixed effects, β is a $q \times 1$ vector of unknown fixed effects, Z is a known $n \times m$ matrix associated with random effects, γ is an $m \times 1$ vector of unknown random effects, and e is an $n \times 1$ vector of random errors. Note that G and R are covariance matrices of γ and e, respectively. Then the variance of y is

$$\text{Var}(y) = V = ZGZ^T + R.$$

If we assume covariance matrices G and R are known, Henderson (1963) showed that the best linear unbiased estimator (BLUE) of β is

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y, \tag{2.2.1}$$

and the best linear unbiased predictor (BLUP) of γ is

$$\hat{\gamma} = GZ^T V^{-1}(y - X\hat{\beta}). \tag{2.2.2}$$

Both two equations depend upon the inverse of V. If the number of observations is large, then the calculation of V^{-1} is computationally intensive. Henderson (1950) offered an alternative way to jointly obtain $\hat{\beta}$ and $\hat{\gamma}$ using his mixed-model equations (MME),

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \times \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}.$$

It has been proved that the solutions for $\hat{\beta}$ and $\hat{\gamma}$ from MME are the BLUE and the BLUP, respectively (Henderson, 1963; Henderson et al., 1959).

BLUE and BLUP are based on known covariance matrices. However, in realistic cases, they are usually unknown. A popular method to estimate variance components of G

and R is called the restricted maximum likelihood (REML) method, and the logarithm of the REML function is given by

$$L(\theta) = -\frac{1}{2}\ln V - \frac{1}{2}\ln|X^T V^{-1} X| - \frac{1}{2}(y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta}),$$

where θ consists of unknown variance components, and

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y.$$

2.2.1 HAT Method

In order to select the best model among several models, the prediction errors are utilized to measure and compare the performance. Suppose n_1 independent individuals $(x_1, y_1), \ldots, (x_{n_1}, y_{n_1})$ are the training data and the regression function fitted by the training data is denoted by $\hat{f}_{n_1}(\cdot)$. Then the resulting models can be applied to predict the response variable or phenotypic measurements at the testing data $x_{n_1+1}, \ldots, x_{n_1+n_2}$. Usually, we use the mean square prediction error (MSPE) to measure the accuracy of prediction. And it can be obtained by

$$\frac{1}{n_2}\sum_{i=1}^{n_2} |y_{n_1+i} - \hat{f}_{n_1}(x_{n_1+i})|^2.$$

If data only consists of n independent observations, using the same data to fit and calculate MSPE generally leads to bad estimate of the prediction error, i.e. the prediction error is much lower than the true prediction error. One popular method to eliminate overfitting phenomenon is the cross-validation (CV) analysis. The general idea of k-fold CV (Picard and Cook, 1984) is that we partition the n observations into k equal-sized parts, where $k \leq n$. $k-1$ parts is considered as our training data and the remaining one part is the testing data. We repeat this process k times in order to utilize every part to

estimate the prediction errors. The overall prediction error is averaged across k MSPEs. A special case, known as the leave-one-out cross-validation (LOOCV), occurs with $k = n$. It was developed by Allen (1971, 1974) to calculate the predicted residual error sum of squares (PRESS), which is the average of n MSPEs drawn from the LOOCV analysis. Moreover, PRESS was proposed to use as a criterion to select the best model among several regression models.

Relative to PRESS, the predictability of a model is used to measure the performance as well. It can be expressed by the squared correlation coefficient between the observed and predicted phenotypic values (Xu et al., 2014). And it is also equal to

$$\mathrm{R}^2 = 1 - \mathrm{PRESS/SS},$$

where the value of PRESS is explained above and the SS is the total sum of squares of the response variable, i.e.

$$\mathrm{SS} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

One limitation of k-fold CV analysis is the random partitioning variation problems. To get rid of it, LOOCV is preferred to compute the prediction errors or predictabilities. However, LOOCV analysis will be computationally expensive when the sample size n is large. Therefore, HAT method was proposed to estimate the prediction errors in order to reduce the computational burden. In general, HAT method is a method to evaluate the prediction errors using the whole sample only once. Cook (1977, 1979) derived the formula to calculate PRESS for OLS by adjusting the coefficients of the linear regression model using the leverage values of observations (the diagonal elements of HAT matrix $H = X(X^TX)^{-1}X^T$). Then Golub et al. (1979) extended the HAT method to the mixed

effects model by finding the optimal shrinkage factor in ridge regression. Four decades later, Gianola and Schön (2016) summarized how to implement the HAT method to compute the prediction errors by running the model only once. This paper derived the HAT methods for linear regression, ridge regression, Bayesian alphabet models, etc. In addition, for the ridge regression, it was claimed that using the shrinkage factor λ estimated by the whole sample instead of using the value estimated within each fold doesn't affect the prediction errors too much, especially for the LOOCV analysis. Xu (2017) subsequently investigated the difference between them, and he provided the derivation of calculation of the LOOCV predictability for a mixed model as well. Here are the details about using HAT method under linear mixed model to obtain the predictability.

Recall a linear mixed model can be written as

$$y = X\beta + Z\gamma + e, \tag{2.2.3}$$

$$\gamma \sim \text{MVN}_m(0, I\sigma_\gamma^2),$$

$$e \sim \text{MVN}_n(0, I\sigma_e^2).$$

Let us define the predicted random effects as

$$\hat{r} = \hat{y} - X\hat{\beta}, \tag{2.2.4}$$

and the observed random effects as

$$r = y - X\hat{\beta},$$

where

$$\hat{y} = X\hat{\beta} + Z\hat{\gamma}.$$

Substituting (2.2.1) and (2.2.2) into (2.2.4), we have

$$\hat{r} = \sigma_\gamma^2 Z Z^T V^{-1} r = Hr,$$

where H is called the HAT matrix. The estimated error vector is

$$\hat{e} = y - \hat{y} = r - \hat{r} = r - Hr = (I - H)r.$$

The predicted error for the i^{th} observation is

$$e_i = y_i - x_i^T \hat{\beta}_{[-i]} = (1 - h_{ii})^{-1} \hat{e}_i,$$

where $\hat{\beta}_{[-i]}$ is the estimate of β with the i^{th} observation deleted and h_{ii} is the i^{th} diagonal element of the HAT matrix, called the leverage value of observation i. Then the PRESS is

$$\text{PRESS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (1 - h_{ii})^{-2} \hat{e}_i^2.$$

Accordingly, the predictability of the model is

$$\text{R}^2 = 1 - \frac{\text{PRESS}}{\text{SS}}$$

where

$$\text{SS} = \sum_{i=1}^{n} (r_i - \bar{r})^2$$

is the total sum of squares of y after correction for the fixed effects.

2.3 Bayesian Alphabet Models

Apart from regularized linear regression models and linear mixed models, Bayesian approaches can also be used in genomic selection and GWAS. The key features of Bayesian approaches compared with previous models are Bayesian approaches incorporate another

information into the models and consider uncertainty of unknown parameters (Zhou et al., 2013). Frequentist inference usually treats model parameters as fixed values, while Bayesian inference treats them as random variables and assigns informative prior beliefs (prior distributions) to them. Then the posterior distributions of unknown parameters can be obtained using Bayes' theorem. Moreover, the posterior distributions describe uncertainty in parameter estimates. In genomic selection and GWAS, Bayesian alphabet models, BayeA, BayesB, and BayesC, are widely used (Meuwissen et al., 2001; Verbyla et al., 2009). The main differences among them are different model assumptions.

2.3.1 BayesA

Linear mixed models assume marker effect sizes come from normal distributions with homogeneous variances if the covariance matrix $R = I\sigma_\gamma^2$. It is not realistic to assume all effect sizes have equal variances. If the variances are varied from marker to marker, then BayeA is derived. The model can be showed in the expression below

$$y = \mu + X\beta + e,$$

$$e|\sigma_e^2 \sim \text{MVN}_n(0, I\sigma_e^2),$$

$$\beta_i|\sigma_i^2 \sim \text{N}(0, \sigma_i^2),$$

where y is an $n \times 1$ vector of a response variable measured on n individuals, μ is an $n \times 1$ vector of intercept terms, X is an $n \times m$ design matrix, β is an $m \times 1$ vector of unknown effect sizes, e is an $n \times 1$ vector of random errors, MVN_n represents the n-dimensional multivariate normal distribution, and $i = 1, \ldots, m$.

Suppose the prior distributions of σ_e^2 and σ_i^2 are

$$\sigma_e^2 \sim \chi^{-2}(-2, 0),$$

and

$$\sigma_i^2 \sim \chi^{-2}(\nu, \tau^2),$$

where ν is the number of degrees of freedom and τ^2 is a scale parameter. Since a scaled inverted chi-square distribution is a conjugate prior, the posterior distributions of σ_e^2 and σ_i^2 can be easily obtained and given by

$$\sigma_e^2|e \sim \chi^{-2}(n-2, e^T e),$$

and

$$\sigma_i^2|\beta_i \sim \chi^{-2}(\nu+1, \tau^2+\beta_i^2).$$

Then the Gibbs sampling algorithm can be applied to estimate effect sizes and variances.

2.3.2 BayesB

BayesB assumes that a large number of markers have no genetic variances and a small number of markers have genetic variances. The model can be expressed by

$$y = \mu + X\beta + e,$$

$$e|\sigma_e^2 \sim \mathrm{MVN}_n(0, I\sigma_e^2),$$

$$\beta_i|\sigma_i^2 \sim \mathrm{N}(0, \sigma_i^2),$$

where y is an $n \times 1$ vector of a response variable measured on n individuals, μ is an $n \times 1$ vector of intercept terms, X is an $n \times m$ design matrix, β is an $m \times 1$ vector of unknown

effect sizes, e is an $n \times 1$ vector of random errors, MVN_n represents the n-dimensional multivariate normal distribution, and $i = 1, \ldots, m$.

Similarly, the prior distribution of σ_e^2 is still the scaled inverted chi-square distribution with parameters $\{-2, 0\}$. But for σ_i^2, it is modified by

$$\sigma_i^2 \sim \chi^{-2}(\nu, \tau^2) \text{ with probability } 1 - \pi,$$

$$\sigma_i^2 = 0 \text{ with probability } \pi,$$

where ν is the number of degrees of freedom, τ^2 is a scale parameter, and π denotes a high density which reflects the proportion of markers without genetic effects. The posterior distributions of σ_e^2 and σ_i^2 remain the scaled inverted chi-squared distributions. However, the Gibbs sampler of BayesA can not be implemented to sample the values of σ_i^2 and β_i here because sampling $\sigma_i^2 = 0$ is impossible when $\beta_i^2 \neq 0$. Thus, we need to take advantage of the joint distribution of σ_i^2 and β_i given y^*, i.e.

$$p(\sigma_i^2, \beta_i | y^*) = p(\beta_i | \sigma_i^2, y^*) \times p(\sigma_i^2 | y^*),$$

where y^* is the response variable y corrected for the overall mean and $\beta_{[-i]}$ (genetic effects without β_i). Then σ_i^2 can be sampled from $p(\sigma_i^2 | y^*)$ and β_i can be sampled from $p(\beta_i | \sigma_i^2, y^*)$.

This algorithm is computationally intensive because the Metropolis-Hasting (MH) algorithm is implemented to sample the values from $p(\sigma_i^2 | y^*)$. Cheng et al. (2015) introduced three different Gibbs samplers to sample the parameters without the MH algorithm. They enhance the running speed twice faster than the Gibbs sampler with the MH algorithm.

2.3.3 BayesC

Since original BayesB is computationally expensive, Verbyla et al. (2009) proposed and suggested a new Bayesian method, which is known as method BayesC. Then model can be described as

$$y = \mu + X\beta + e,$$

$$e|\sigma_e^2 \sim \text{MVN}_n(0, I\sigma_e^2),$$

$$\beta_i|u_i, \sigma_i^2 \sim u_i\text{N}(0, \sigma_i^2/100) + (1 - u_i)\text{N}(0, \sigma_i^2),$$

where y is an $n \times 1$ vector of a response variable measured on n individuals, μ is an $n \times 1$ vector of intercept terms, X is an $n \times m$ design matrix, β is an $m \times 1$ vector of unknown effect sizes, e is an $n \times 1$ vector of random errors, u_i is a binary indicator variable for the i^{th} marker (u_i=$\{0,1\}$), MVN_n represents the n-dimensional multivariate normal distribution, and $i = 1, \ldots, m$.

The prior distributions of σ_i^2 and u_i are

$$\sigma_i^2 \sim \chi^{-2}(\nu, \tau^2),$$

and

$$u_i \sim \text{Bernoulli}(\pi_i),$$

where ν is the number of degrees of freedom, τ^2 is a scale parameter and π_i denotes a high density. Based on the hierarchical prior assumptions, marker effects can be sampled from a mixture of the scaled t distributions. The indicator variable u_i is sampled from the posterior distribution $p(u_i = 1|\beta_j, \sigma_i^2, u_{[-i]}, y)$, i.e.

$$\text{Bernoulli}(\frac{p(\beta_j|u_{[-i]}, u_i = 1)(1 - \pi_i)}{p(\beta_j|u_{[-i]}, u_i = 1)(1 - \pi_i) + p(\beta_j|u_{[-i]}, u_i = 0)\pi_i}),$$

where $u_{[-i]}$ is all the indicator variables with u_i deleted.

2.4 Predictions

Suppose n_1 independent individuals $(x_1, y_{11}), \ldots, (x_{n_1}, y_{1n_1})$ are the training data and n_2 independent individuals $x_{n_1+1}, \ldots, x_{n_1+n_2}$ are the testing data. In this section, let us discuss how we can calculate the predictions of a quantitative trait based on the model fitted in the training data, which is denoted by $\hat{f}_{n_1}(\cdot)$. According to previous sections, we know that the basic models for regularized linear regression models and Bayesian alphabet models are the same, i.e.

$$y = \mu + X\beta + e.$$

The differences among them are effect size assumptions and parameter estimates. Therefore, for these two types of models, the predictions at the testing data, $\hat{y}_2$, can be easily calculated by

$$\hat{y}_2 = \hat{f}_{n_1}(X_2) = \hat{\mu} + X_2\hat{\beta},$$

where $X_2 = (x_{n_1+1}, \ldots, x_{n_1+n_2})^T$.

However, the predictions in the linear mixed model is not straightforward. Here are the details. Recall the linear mixed model can be formulated as equation (2.2.3). Define the kinship matrix as $K = ZZ^T$. Then the extended kinship matrix is

$$\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22}, \end{bmatrix}.$$

Here K_{11} is the $n_1 \times n_1$ kinship matrix with respect to the training observations, K_{22} is the $n_2 \times n_2$ kinship matrix with respect to future crosses, and K_{12} or K_{21} is the relationship

matrix with respect to current crosses and future crosses. Then the predicted breeding values of y_2 is

$$\hat{y}_2 = X_2\hat{\beta} + \sigma_\gamma^2 K_{21} V_1^{-1}(y_1 - X_1\hat{\beta}),$$

where

$$\hat{\beta} = (X_1^T V_1^{-1} X_1)^{-1} X_1^T V_1^{-1} y_1,$$

$$V_1 = \sigma_\gamma^2 K_{11} + I\sigma_e^2,$$

X_1 is an $n_1 \times q$ design matrix, X_2 is an $n_2 \times q$ design matrix, and y_1 is an $n_1 \times 1$ vector associated with current quantitative traits.

2.5 Summaries

All the models I discussed above belong to polygenic modeling, which is regressing genotypic values on a quantitative trait and estimating genetic effect sizes simultaneously. Table 2.1 lists the summaries of effect size distributions that have been developed to deal with high dimensionality of predictors. Regularized linear model estimates are equivalent to Bayesian estimates when regression coefficients are assigned appropriate priors. For instants, Bayesian estimates are identical to LASSO estimates when σ_i is assigned a prior distribution, an exponential prior. Then the marginal prior distribution of β_i is double exponential. Ridge regression is identical to BLUP if we define the shrinkage factor is the variance ratio in (2.2.3), i.e. $\lambda = \sigma_e^2/\sigma_\gamma^2$. One thing may be confused about table 2.1 is effect size distributions for Bayesian alphabet models. According to model assumptions, effect size distributions of Bayesian alphabet models are related to normal distributions. Since the prior distribution of genetic variance, σ_i^2, is the scaled inverted chi-square distribution, the

marginal prior of β_i results in the scaled t distribution. That is the reason that prior distributions of Bayesian alphabet models are associated with the scaled t distributions.

Different distributions of effect sizes lead to different model behaviors. Figure 2.1 (de los Campos et al., 2013) illustrates distributions of effect sizes listed in table 2.1. Relative to normal priors, scaled t and double exponential are known as thick-tail priors. A thick-tail prior or a scaled t mixture prior has a property that it tends to strongly shrink small effect sizes to zero and lightly shrink large effect sizes compared with a normal prior. The scaled point-t distribution makes variable selection feasible because of a point mass at zero with a high density.

Table 2.1 Effect size (β_i) distributions, their formulas and model names. s denotes the scale parameter. $d.f.$ denotes the degrees of freedom parameter. δ_0 denotes a point mass at zero. π denotes a high probability value.

Effect size distribution	Formula	Model name
Normal	$\beta_i \sim \mathrm{N}(0, \sigma^2)$	BLUP, Ridge
Scaled t	$\beta_i \sim \mathrm{t}(d.f., s)$	BayesA
Scaled point-t	$\beta_i \sim (1-\pi)\mathrm{t}(d.f., s) + \pi\delta_0$	BayesB
Scaled t mixture	$\beta_i \sim (1-\pi)\mathrm{t}(d.f._1, s_1) + \pi\mathrm{t}(d.f._2, s_2)$	BayesC
Double exponential	$\beta_i \sim \mathrm{DE}(s)$	LASSO

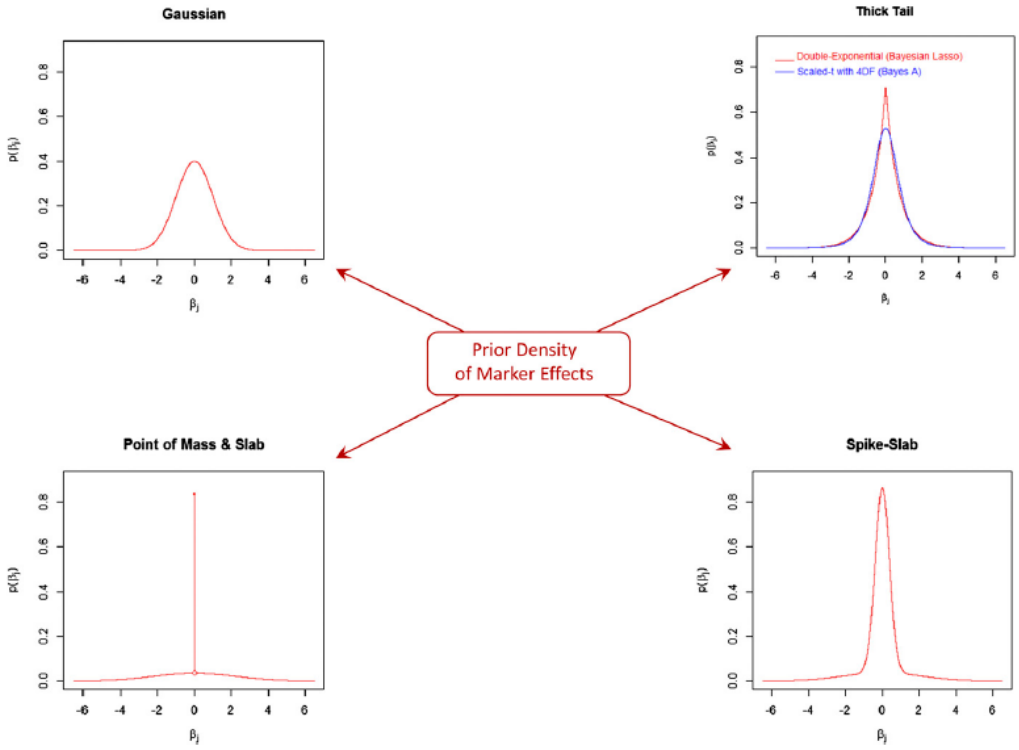

Figure 2.1 Prior densities of marker effects. Top left panel is a normal prior. Top right panel is thick-tail priors. Bottom right panel is a scaled t mixture prior. Bottom left panel is a scaled point-t panel.

Chapter 3

A Stochastic Restricted Maximum Likelihood Method

3.1 Introduction

In the past, plant breeders had to spend a lot of years to obtain desirable hybrid crosses by planting various hybrid rice varieties. Recently a new approach has been developed to accelerate the breeding process, i.e. genomic selection. Genomic selection is a methodology that provides desirable candidates with a shorter breeding cycle using millions of molecular markers information to predict future individuals. Therefore, the predicted phenotypic values evaluated by the statistical models can easily detect the desirable individuals to reduce the cost of traditional breeding. This methodology has been used in many areas, such as humans, animals, and plants, however, it started to be applied in the hybrid prediction after 2014 (Xu et al., 2014).

Genomic selection is a marker-assisted selection method and numerous single-nucleotide polymorphisms (SNPs) are involved in it, therefore, it is important to implement the efficient and effective models to accurately predict phenotypic traits. Modern techniques facilitate the utilization of complicated linear or nonlinear statistical models for the phenotypic predictions. Even through nonlinear models can achieve higher accuracy sometimes, linear models are easy to interpret and make marker selection feasible. Due to it, this dissertation only focuses on the linear models. The multiple linear regression model is one of the simplest and basic linear models. However, in plant breeding it is hard to implement it because of high dimensionality and multicollinearity of markers (Crossa et al., 2017). High dimensionality of markers means the number of markers substantially exceeds the number of individuals, which leads to nonexistence of design matrix inverses. And multicollinearity of markers results in the wrong signs and insignificance of coefficients. In order to address those problems, regularized linear regression models, such as ridge regression (Hoerl and Kennard, 1970a,b) and least absolute shrinkage selection operator (LASSO) regression (Tibshirani, 1996), the linear mixed model (Henderson, 1975), and Bayesian alphabet models, such as BayesA, BayesB and BayesC approaches (Meuwissen et al., 2001; Verbyla et al., 2009), have been proposed and widely used. Regularized linear models estimate unknown parameters by minimizing the least square and penalty terms. The regression coefficients influenced by different penalty terms are distinguishing, where shrinkage of estimates or variable selection or even both can be achieved. For example, ridge regression uses l_2 penalty term to shrink coefficients while LASSO regression uses l_1 penalty term to shrink coefficients and select variables. In contrast, the linear mixed model, also known as the best linear unbiased prediction (BLUP) model, assigns a prior distribution with homogeneous variances to all

markers instead of adding penalty terms into the loss function. If heterogeneous variances are adopted across all markers, i.e. each marker has an unequal effect on phenotypic values, and each variance is assigned a prior distribution, then method BayesA is applied. Relative to method BayesA, if only a small number of markers have heterogeneous variances and remaining markers have null effects on the quantitative trait values, then method BayesB is adopted. Based on the definition of method BayesB, it is obvious that it makes marker selection feasible. If we assign ignorable values of variance to the remaining markers instead of treating them as zero, then method BayesC is derived.

Apart from genomic selection, in genetics, a genome-wide association study (GWAS) (Risch and Merikangas, 1996) is another crucial field, which is a study of exploiting the associations between some specific SNPs and valuable phenotypic traits using the entire genome information. For example, in plant breeding, we can detect which SNPs have sizable effects on an important trait, such as yield or grain. One of the most popular methods proposed to detect significant markers is linear mixed model (LMM) (Yu et al., 2006). However, this method is computationally costly because of large and many matrix multiplications and inverses. In order to speed up the running time, several algorithms have been developed, such as efficient mixed-model association (EMMA) (Kang et al., 2008), genome-wide efficient mixed-model association (GEMMA) (Zhou and Stephens, 2012), etc. The methods listed above test a single marker at a time, which means all markers need to be scanned to identify the relationships with the quantitative trait. One limitation of these approaches is the multiple testing problem, that is if a high-density SNP array is employed, then the level of significance should be extremely stringent. Further, the single-SNP approaches may not detect any single significant marker if markers are highly correlated to each other. In

terms of these two limitations, the methods estimating marker coefficients simultaneously outperform the single-SNP approaches. Among the methods I introduced above, LASSO and BayesB belongs to variable selection methods, or known as selective shrinkage methods, which can be used in GWAS as well. It is very convenient to use both models in R. LASSO can be easily implemented using GlmNet/R program (Friedman et al., 2010), and BayesB can be easily implemented using BGLR/R program (Pérez and de Los Campos, 2014). As mentioned earlier, LASSO uses l_1 penalty term to shrink the coefficients of some markers to exactly zero, but BayesB assumes the variance of each marker is zero with a known high probability or is a nonzero value with a low probability. Theoretically, BayesB can delete markers with zero variance.

It has been showed that BLUP is more robust than LASSO and BayesB under some conditions (Xu et al., 2014), and that is the motivation of proposing a new approach based on BLUP in genomic selection and GWAS. The contribution of the study is taking into account of the idea of a stochastic EM algorithm and a small group of markers with additional variances which facilitates to select variables. The general idea of the proposed approach is dividing markers into two classes, i.e. the large effect class and the small effect class. More details of it are described in the next section.

3.2 Materials and Methods

3.2.1 A Stochastic Restricted Maximum Likelihood Model

This approach consists of the REML step and the stochastic step. In the REML step, variance components can be estimated given the cluster labels of markers. In the

stochastic step, the cluster labels of markers are updated given variance components. The details of this approach are described as follows.

Consider m loci that are divided into m_S loci with small effects and m_L loci with large effects where $m_S + m_L = m$. Then define the indicator variable or cluster label of locus i by

$$u_i = \begin{cases} 0 & \text{if marker } i \text{ belongs to the large effect cluster} \\ 1 & \text{if marker } i \text{ belongs to the small effect cluster} \end{cases}$$

where $u_i \sim \text{Bernoulli}(\pi)$ is a Bernoulli variable with a predetermined high probability $\pi = 0.95$. This prior distribution assumes that 95% of the loci have small effects on the phenotypic values and the remaining 5% markers have sizable effects. It is possible to assign a conjugate prior, a Beta distribution, to π so that the value of π can be updated using the data information, but for simplicity of the method we set a constant value for this parameter at first.

A linear mixed model with two random components can be formulated as

$$y = X\beta + Z_S\gamma_S + Z_L\gamma_L + e, \tag{3.2.1}$$

$$\gamma_S \sim \text{MVN}_{m_S}(0, I\sigma_S^2),$$

$$\gamma_L \sim \text{MVN}_{m_L}(0, I\sigma_L^2),$$

$$e \sim \text{MVN}_n(0, I\sigma_e^2).$$

Here y is an $n \times 1$ observation vector, X is an $n \times q$ covariate matrix including an intercept vector, β represents a $q \times 1$ non-genetic fixed effect vector, Z_S and Z_L are $n \times m_S$ small effect marker and $n \times m_L$ large effect marker matrices, γ_S and γ_L are $m_S \times 1$ and $m_L \times 1$ genetic effect vectors contributed by the small effect loci and the large effect loci, respectively, e is

an $n \times 1$ vector of error terms. Note that MVN_a represents the a-dimensional multivariate normal distribution, and $\sigma_S^2 < \sigma_L^2$. Define Z as the whole marker matrix and Z_i as the numerical coded values of the i^{th} marker for all individuals. Then the two polygenic vector which are random effect can be written as

$$Z_S\gamma_S = \sum_{i=1}^{m} u_i Z_i \gamma_i,$$

and

$$Z_L\gamma_L = \sum_{i=1}^{m} (1-u_i) Z_i \gamma_i,$$

where γ_i is the effect of the i^{th} locus and is treated as a random effect with mean zero and variance

$$\text{Var}(\gamma_i) = u_i\sigma_S^2 + (1-u_i)\sigma_L^2,$$

which implies the variance of γ_i is equal to σ_S^2 if the i^{th} locus belongs to the small effect cluster and σ_L^2 if the i^{th} locus belongs to the large effect cluster. Then the expectation of y is

$$\text{E}(y) = X\beta,$$

and the variance V is

$$\begin{aligned}\text{Var}(y) &= Z_S\text{Var}(\gamma_S)Z_S^T + Z_L\text{Var}(\gamma_L)Z_L^T + \text{Var}(e) \\ &= K_S\sigma_S^2 + K_L\sigma_L^2 + I\sigma_e^2.\end{aligned}$$

Here $K_S = Z_S{Z_S}^T$ and $K_L = Z_L{Z_L}^T$ are called kinship matrices.

Given u_i, the unknown parameters $\theta = \{\sigma_S^2, \sigma_L^2, \sigma_e^2\}$ can be estimated using the restricted maximum likelihood (REML) method. After the unknown parameters are estimated, the new labels for each locus can be updated. Bayes' theorem is applied to

sample a new $u_i^{(t+1)}$ conditional on the current value $u_i^{(t)}$ and all other parameter values. Given the current value $u_i^{(t)}$ and $\theta^{(t)}$, the conditional posterior probability of $u_i^{(t+1)}$ is $P(u_i^{(t+1)} = 1|u_i^{(t)}, \theta^{(t)}) = \rho_i$, and ρ_i is expressed as

$$\rho_i = \frac{\pi}{\pi + (1-\pi)exp(\psi)},$$

where

$$\psi = (1 - 2u_i^{(t)})[L(\theta^{(t)}) - L_i(\theta^{(t)})],$$

and $L(\theta)$ is the restricted log likelihood function without changing the labels, and $L_i(\theta)$ is the restricted log likelihood function when the i^{th} marker switches its current cluster label, i.e., from the small effect cluster to the large effect cluster or vice versa, depending on its current position. Therefore, the conditional posterior distribution of u_i can be obtained according to Bernoulli(ρ_i) and a new $u_i^{(t+1)}$ is sampled based on it.

To switch the cluster label for marker i, we need to update the kinship matrices. If the i^{th} marker is currently in the small effect cluster, a switch will place this marker to the large effect cluster and the new variance matrix will be

$$V_{+i} = V - Z_i Z_i^T \sigma_S^2 + Z_i Z_i^T \sigma_L^2.$$

A switch from the large effect cluster to the small effect cluster will result in a new variance matrix of

$$V_{-i} = V - Z_i Z_i^T \sigma_L^2 + Z_i Z_i^T \sigma_S^2.$$

Since the current cluster label for marker i is $u_i^{(t)}$, if $u_i^{(t)} = 1$, the i^{th} marker is currently in the small effect cluster and a switch means placing it to the large effect cluster and thus we should take V_{+i}. Similarly, if $u_i^{(t)} = 0$, we should switch marker i from the large effect

cluster to the small effect cluster and thus V_{-i} should be taken. Incorporating $u_i^{(t)}$ into the establishment of the new variance matrix, we have

$$\begin{aligned} V_i &= V - (1 - 2u_i^{(t)})(Z_i Z_i^T \sigma_L^2 - Z_i Z_i^T \sigma_S^2) \\ &= V - (1 - 2u_i^{(t)})(\sigma_L^2 - \sigma_S^2) Z_i Z_i^T \\ &= V - (1 - 2u_i^{(t)}) \Delta Z_i Z_i^T, \end{aligned}$$

where $\Delta = \sigma_L^2 - \sigma_S^2$ is the difference between the large variance and the small variance. Then the restricted log likelihood function of switching marker i is

$$L_i(\theta) = -\frac{1}{2}\ln|V_i| - \frac{1}{2}\ln|X^T V_i^{-1} X| - \frac{1}{2}(y - X\hat{\beta})^T V_i^{-1}(y - X\hat{\beta})$$

where

$$\hat{\beta} = (X^T V_i^{-1} X)^{-1} X^T V_i^{-1} y.$$

Looping over all markers represents an extremely large computational burden due to calculations of V_i^{-1} and $|V_i|$. However, we can avoid it by taking advantage of Woodbury matrix identities for the inverse and the matrix determinant lemma for the determinant, which are

$$\begin{aligned} V_i^{-1} &= [V - (1 - 2u_i^{(t)}) \Delta Z_i Z_i^T]^{-1} \\ &= V^{-1} - V^{-1} Z_i [Z_i^T V^{-1} Z_i - (1 - 2u_i^{(t)}) \Delta^{-1}]^{-1} Z_i^T V^{-1} \end{aligned}$$

for the inverse and

$$\begin{aligned} |V_i| &= |V - (1 - 2u_i^{(t)}) \Delta Z_i Z_i^T| \\ &= |V||2u_i^{(t)} \Delta Z_i^T V^{-1} Z_i| \end{aligned}$$

for the determinant. Therefore, the obvious advantage computationally is that we do not need to calculate the marker specific matrix inverse and determinant anew but simply update from V^{-1} and $|V|$. Looking back at the likelihood function, we notice that V_i^{-1} never occurs alone but always exists in a quadratic form like $a^T V_i^{-1} b$, where a and b can be X, Z_i, y or $y - X\hat{\beta}$. These quadratic forms are expressed by

$$a^T V_i^{-1} b = a^T V^{-1} b - a^T V^{-1} Z_i [Z_i^T V^{-1} Z_i - (1 - 2u_i^{(t)})\Delta^{-1}]^{-1} Z_i^T V^{-1} b.$$

Tremendous computational time can be saved via this special algorithm.

The proposed method involves stochastic sampling for the cluster labels of markers given the variance component parameters and REML estimation of variance parameters given the cluster labels of markers. Iterations are required over the stochastic step and the REML step, and thus the method is called stochastic restricted maximum likelihood (SREML) method. During the SREML process, for each iteration the predictability is evaluated using the HAT method (Xu, 2017), where the leave-one-out cross validation (LOOCV) mean squared prediction error (MSPE) can be obtained by fitting the data values only once. Then the maximum predictability and the corresponding u vector will be recorded. The number of iterations is quite arbitrary but does not need to be very large. If 20 consecutive iterations fail to identify better performance, we may stop the SREML process and report the predictability and its associated u vector. The longer the iterations, the higher the chance to identify the best classification.

An enhancement of the SREML is to estimate π from the data rather than to fix its value at 0.95. Suppose we assign a conjugate prior to π, Bata(a, b). Then the posterior distribution remains Beta, i.e. Bata$(a + m_S, b + m_L)$, where $m_S = \sum_{i=1}^{m} u_i$ and

$m_L = \sum_{i=1}^{m}(1 - u_i)$. One of the reasonable choices of a and b is $a = m$ and $b = 0.1m$. Figure 3.1 and figure 3.2 illustrate Beta distributions when m=5,000 and m=100,000. The comparison of two figures tell us no matter the value of m is small or large, a random variable sampled from a Beta distribution will be approximately equal to 0.91. It implies we expect a large number of loci to be small effect cluster. From the posterior distribution, Bata$(a + m_S, b + m_L)$, a new π is sampled. $\pi^{(t+1)}$ is then used to update the u vector.

Figure 3.3 is an example of the SREML procedure when we adopt the HAT method as our criterion and it shows the predictability against the iteration number. The red and blue dashed lines represent the iteration number is 50 and 160 respectively; the red and blue dots represent the maximum predictability locations for different iteration numbers. This figure illustrates that the procedure of the SREML randomly goes up and down, and we have more chance to obtain a better result as the iteration number increases. We also can conclude that the process is stochastically around an invisibly horizontal line and the best predictability will be found during the process.

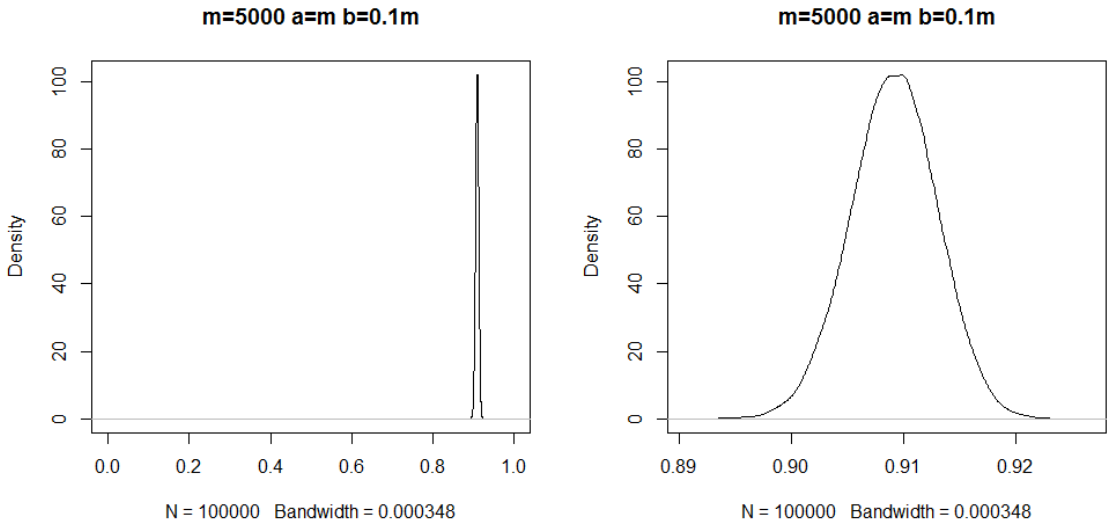

Figure 3.1 Left panel is the density plot of Beta distribution with a=5,000 and b=500. Right panel enlarges the peak part of the left panel.

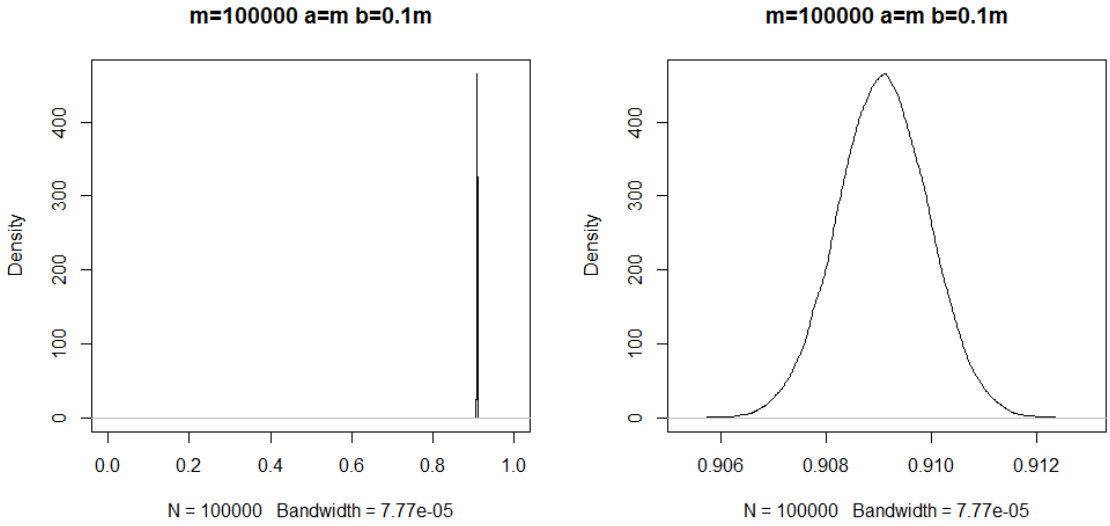

Figure 3.2 Left panel is the density plot of Beta distribution with a=100,000 and b=10,000. Right panel enlarges the peak part of the left panel.

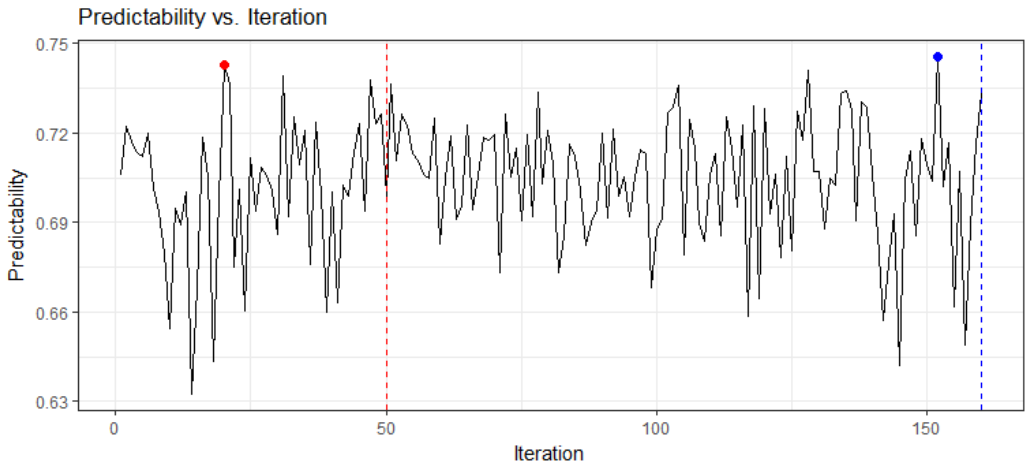

Figure 3.3 Predictability of a trait plotted against the iteration numbers. The red dash line represents the iteration number is 50, and the blue dash line represents the iteration number is 160. Two solid points are the corresponding iteration numbers with highest predictabilities.

3.2.2 Algorithm

Start at $t = 0$ with a randomly initiated value $u_i^{(0)}$, $i = 1, 2, ..., m$ and $\pi^{(0)} = 0.95$. Given $u_i^{(0)}$ and $\pi^{(0)}$, the algorithm generates estimates as follows:

Step 1 (REML step): Estimate variance components $\{\sigma_S^2, \sigma_L^2, \sigma_e^2\}$ by the restricted log likelihood function given the current estimates $\{u_i^{(t)}, \pi^{(t)}\}$.

Step 2 (Stochastic step): Update the cluster label $u_i^{(t+1)}$ and $\pi^{(t+1)}$ given the estimated variance components $\{\sigma_S^2, \sigma_L^2, \sigma_e^2\}$.

Step 3: Iterate between step 1 and step 2 until the maximum number of iterations has been reached.

Step 4: Select the estimates of the variance components and the u vector associated with the maximum predictability.

3.3 Results

3.3.1 Genomic Selection

Some statistical values, such as MSE, are commonly computed to compare the performance among prediction models. However, it will cause overly optimistic estimates of prediction error if the training data set is used to fit models and calculate prediction errors simultaneously. Therefore, cross-validation (CV) method is widely used to compare statistical models for eliminating overfitting phenomenon(Meuwissen et al., 2001). The general idea of k-fold CV (Picard and Cook, 1984) is to partition n observations into k approximately equal-sized parts. Observations of k-1 folds are considered as the training data to build prediction models and the remaining fold is the testing data set to estimate the prediction errors. This process will be repeated k times such that every fold will be the testing data set, and the overall prediction error is the average of k testing MSEs. Figure 3.4 is an example of 5-fold CV. A data set is divided into 5 folds and each fold will be our testing data.

Moreover, to straightforwardly compare the results among different models and different data types, the predictability, which is approximately equal to squared correlation between predicted response variable and observed response variable (Xu et al., 2016), will be the criterion in this section, where the range is from 0 to 1.

The sources of the data sets applied in this study are publicly available and have been published online. And they can be obtained from the literatures cited.

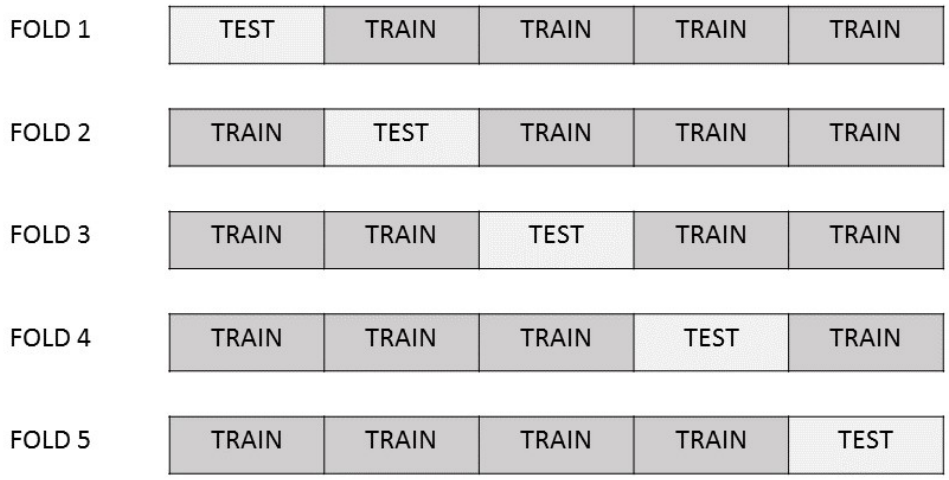

Figure 3.4 This is an example to depict 5-fold CV. A data set is divided into 5 folds and each fold will be our testing data.

3.3.1.1 A Simulation Study

A simulation study was performed to demonstrate the comparison of the predictability under six methods, which are SREML, BLUP, LASSO, BayesA, BayesB, and BayesC. LASSO results were obtained via GlmNet in R package. And Bayesian alphabet models implemented BGLR in R package with the number of iterations=3,500 and the number of burn-in=500. The real 1,619 genetic markers were used to generate 278 observations (More details about these real genetic markers are described in real-world data analysis). And simulated data y was generated using the model 3.2.1, where $q = 1$ and the fixed effect term only contained the intercept term μ, whose value was fixed to a constant value 100. There were two scenarios in this simulation study: in Scenario I, only one genetic marker was randomly selected with large genetic effect ($\gamma_L = 20$), the remaining genetic effects were sampled from $\mathrm{N}(0, I\sigma_S^2)$ and the error term was sampled from $\mathrm{N}(0, I\sigma_e^2)$; in Scenario

II, 5% of 1,619 genetic markers were randomly selected and their effects were sampled from $\mathrm{N}(0, I\sigma_L^2)$, the remaining genetic effects were sampled from $\mathrm{N}(0, I\sigma_S^2)$ and the error term was sampled from $\mathrm{N}(0, I\sigma_e^2)$. The average value of ten 10-fold CV results was calculated to measure the predictive performance since k-fold CV would lead to the random partitioning variation problem. Note that large effect marker positions were consistent in ten 10-fold CVs for either scenario.

In Chapter 2, I introduce the formulas to calculate the predictive breeding values of BLUP method for a breeding population (testing data). But it only works for BLUP with one random component. Here let us discuss the prediction of BLUP with two random components first. Define the dimension of observed individuals as n_1, the dimension of future individuals as n_2, and the dimension of genetic markers for one individual as m. Then the extended kinship matrices can be expressed as

$$\begin{bmatrix} K_{S11} & K_{S12} \\ K_{S21} & K_{S22}, \end{bmatrix},$$

and

$$\begin{bmatrix} K_{L11} & K_{L12} \\ K_{L21} & K_{L22}, \end{bmatrix}.$$

Here K_{S11} is an $n_1 \times n_1$ kinship matrix with respect to the training observations measured on small effect markers, K_{L11} is an $n_1 \times n_1$ kinship matrix with respect to the training observations measured on large effect markers, K_{S22} is an $n_2 \times n_2$ kinship matrix with respect to future crosses measured on small effect markers, K_{L22} is an $n_2 \times n_2$ kinship matrix with respect to future crosses measured on large effect markers, and (K_{S12}, K_{S21}) and (K_{L12}, K_{L21}) are the relationship matrices with respect to current and future crosses

measured on small effect markers and large effect markers, respectively. Note that we don't need to calculate all components for prediction, and we only need K_{S11}, K_{L11}, K_{S21} and K_{L21}. Then the predicted breeding values of y_2 is

$$\hat{y}_2 = X_2\hat{\beta} + (\sigma_S^2 K_{S21} + \sigma_L^2 K_{L21})V_1^{-1}(y_1 - X_1\hat{\beta}),$$

where

$$\hat{\beta} = (X_1^T V_1^{-1} X_1)^{-1} X_1^T V_1^{-1} y_1,$$

$$V_1 = \sigma_S^2 K_{S11} + \sigma_L^2 K_{L11} + I\sigma_e^2,$$

X_1 is an $n_1 \times q$ design matrix, X_2 is an $n_2 \times q$ design matrix, and y_1 is an $n_1 \times 1$ vector associated with current phenotypic values.

As mentioned earlier, the predictability is measured using $\text{corr}^2(y_2, \hat{y}_2)$, where corr denotes the correlation coefficient. If we set $\sigma_S^2 = 0.04$, $\sigma_L^2 = 400$, $\sigma_e^2 = 1$, the initial value of π is 0.95 and the iteration number of SREML is 50, then the predictabilities are displayed in table 3.1. All methods produces similar results for both scenarios. Among these six methods, SREML slightly outperforms other methods. BayesC has the worst results, the predictabilities are 0.9261 and 0.9577 for Scenario I and II.

Table 3.2 lists the estimated large marker effect under six methods in Scenario I. The parameters were estimated by the whole sample. It shows BLUP heavily shrinks the coefficients to zero. Estimates from LASSO, BayesA, BayesB are close to the true value of marker effects. Figure 3.5 illustrates true large marker effects and the estimated marker effects under six models in Scenario II. Figure 3.6 illustrates true small marker effects and

Table 3.1 Comparison of the predictability for simulated data under six methods. All predictabilities were calculated by squared correlation between predicted response variable and observed response variable. These averaged results were drawn from ten different 10-fold CVs to reduce the partitioning variation. All methods were evaluated under the same fold IDs. The iteration number of SREML was 50. Scenario I represents only one marker had a constant large effect (20), and remaining markers had small effects and were randomly sampled from normal distribution with mean 0 and variance 0.04. Scenario II represents 81 markers (5% of 1619 markers) had large effects and again remaining markers were drawn from variance 0.04.

Scenario	SREML	BLUP	LASSO	BayesA	BayesB	BayesC
I	0.9859	0.9793	0.9781	0.9850	0.9836	0.9261
II	0.9967	0.9950	0.9776	0.9764	0.9639	0.9577

Table 3.2 The estimated large marker effects under six methods in Scenario I.

Marker	β_{true}	β_{SREML}	β_{BLUP}	β_{LASSO}	β_{BayesA}	β_{BayesB}	β_{BayesC}
502	20	7.56	1.94	17.59	20.61	20.56	8.47

the estimated marker effects under six models in Scenario II. Variance components estimated by SREML and BLUP in Scenario II are listed in Appendix 3.B.

3.3.1.2 Real-world Data Analysis

The IMF2 population (Xu et al., 2016), i.e. the immortalized F_2 population, was used to demonstrate the comparison of predictabilities under six methods applied in the simulation study. A total of 278 crosses were generated by randomly pairing the 210 recombinant inbred lines (RILs), and a total of 1,619 bins inferred from 270,820 SNPs were treated as the genetic markers (Yu et al., 2011). Four traits were analyzed to examine the prediction accuracy, which were tiller number per plant (TILLER), grain number per plant

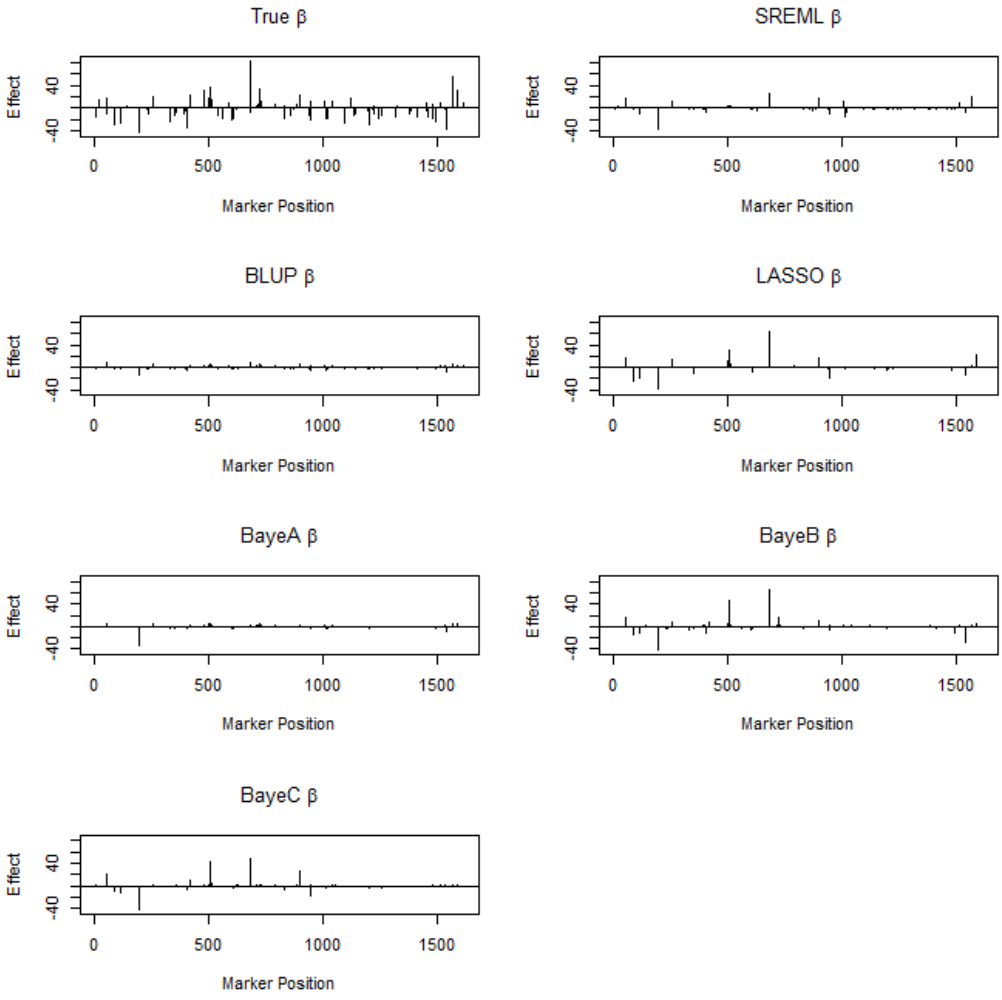

Figure 3.5 The true large marker effects and the estimated large marker effects under six methods in Scenario II.

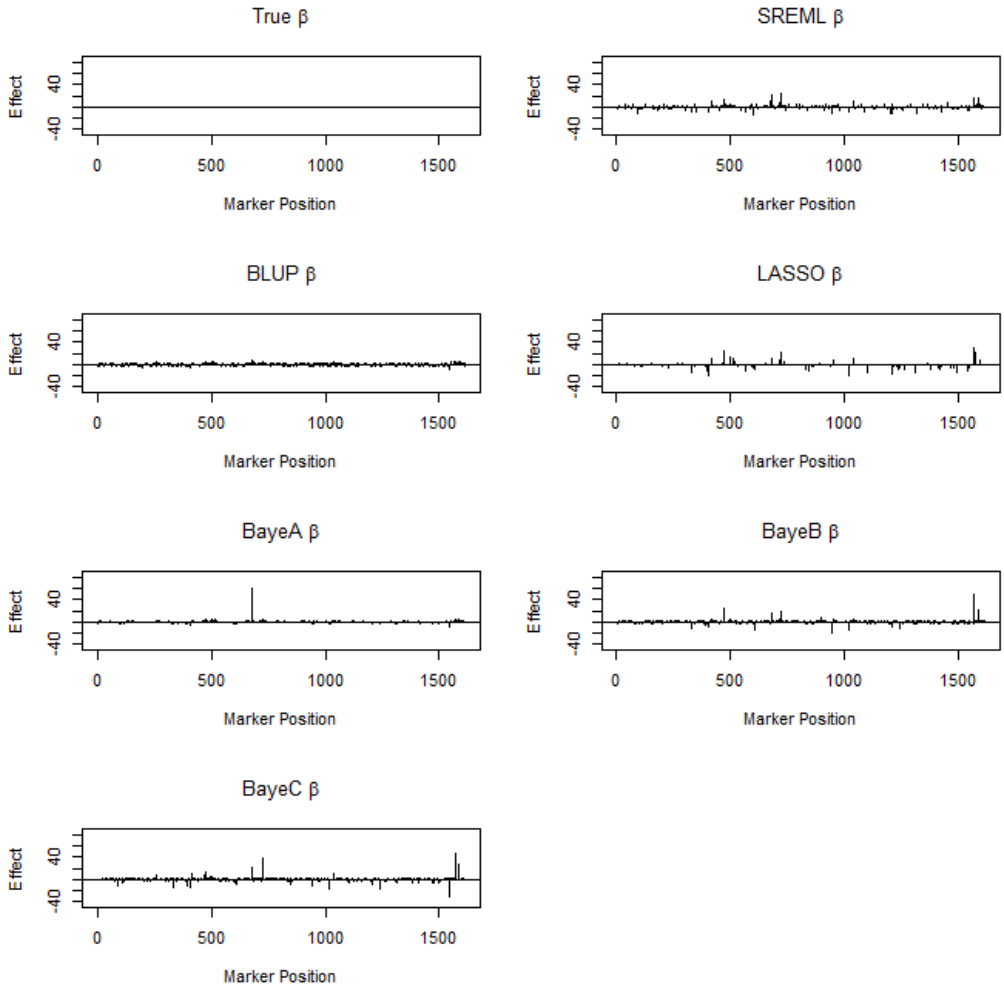

Figure 3.6 The true small marker effects and the estimated small marker effects under six methods in Scenario II.

(GRAIN), yield (YIELD), and 1000-grain weight (KGW), respectively. The numerical value for the k^{th} marker in the j^{th} cross was coded as

$$z_{jk} = \begin{cases} -1 & \text{for aa} \\ 0 & \text{for Aa} \\ +1 & \text{for AA} \end{cases}$$

where aa, Aa and AA were three genotypes of a single marker. Four traits showed in Table 3.3 are predictabilities drawn from ten 10-fold CVs with six methods for this data set. The iteration number of SREML was 50. Compared with the simulation study, SREML results lose its edge. Table 3.3 implies SREML has similar results with other methods, and there is not a unique method outperforming the remaining methods. The model performance depends on the traits we analyzed.

Table 3.3 Comparison of the predictability for the IMF2 population under six methods. All predictabilities were calculated by squared correlation between predicted response variable and observed response variable. The averaged results were drawn from ten different 10-fold CVs to reduce the partitioning variation for the IMF2 population. All methods were evaluated under the same fold IDs. The iteration number of SREML was 50.

Trait	SREML	BLUP	LASSO	BayesA	BayesB	BayesC
TILLER	0.2194	0.2374	0.1960	0.2470	0.2452	0.2464
GRAIN	0.3607	0.3659	0.3680	0.3776	0.3826	0.3688
YIELD	0.1411	0.1314	0.1537	0.1401	0.1449	0.1387
KGW	0.6871	0.6958	0.6968	0.7001	0.7045	0.6934

The other data used to compare the prediction accuracy among six method was from 1,495 elite hybrid rice varieties (Huang et al., 2015). A total of 38 agronomic traits were investigated, and the hybrid varieties were planted at two locations, Sanya and Hangzhou in China. The genotypes of the rice hybrids consisted of totally 182,010 SNPs and the

numerical coded value for each marker was the same as the IMF2 population. Several quantitative traits were contained in this data set, such as YIELD, disease-resistance traits, etc. Because YIELD is a very important trait for the hybrid rice and it relates to global food supply which needs to satisfy the increasing human demanding, the YIELDs at two different locations were selected to be analyzed. Table 3.4 indicates the predictabilities among six methods for YIELDs planted in Hangzhou and Sanya. Due to the high quantity of genetic markers, a 5-fold CV was used to evaluate the predictabilities, and the iteration number of SREML was 20. As with the IMF2 data set, the performance of SREML in table 3.4 is moderate, and Bayesian alphabet models barely outperform other models.

Table 3.4 Comparison of the predictability for the elite hybrid rice data under six methods. All predictabilities were calculated by squared correlation between predicted response variable and observed response variable. The averaged results were drawn from a 5-fold CV. All methods were evaluated under the same fold IDs. The iteration number of SREML was 20. Subscript HZ represents the hybrid varieties were planted in Hangzhou. Subscript SY represents the hybrid varieties were planted in Sanya.

Trait	SREML	BLUP	LASSO	BayesA	BayesB	BayesC
YIELD_{HZ}	0.1225	0.1352	0.1219	0.1405	0.1261	0.1358
YIELD_{SY}	0.0612	0.0614	0.0548	0.0678	0.0517	0.0705

3.3.2 GWAS

3.3.2.1 A Simulation Study

Besides genomic selection, the proposed approach can be used to identify quantitative trait loci (QTL). As mentioned earlier, some of variable selection models, i.e. LASSO and BayesB, are able to be applied in GWAS. Hence, these two models were applied in our

simulation study. However, it has been showed that LASSO does not take into account the linkage disequilibrium (LD), i.e. the correlation of covariates in statistics. Therefore, LASSO tends to just select one or few genetic markers if a group of markers are highly correlated with each other (Zou and Hastie, 2005). In the simulation study, results verified it. BayesB approach was also attempted to use to identify QTL. We expect BayesB outperforms LASSO because Bayesian approaches are not significantly implicated by collinearity. Its outputs were not as the same as the outputs from LASSO. All marker coefficients were nonzero values since it is impossible to sample a value from a zero variance. Instead, we used posterior probabilities to determine which markers were included in the large effect class. To achieve stringent variable selection, two options were added into the model when we implemented BGLR in R, i.e. probIn (the probability that a marker enters the model)=0.05 and counts=10^6. The original options were set to new values, the number of iterations=5,000 and the number of burn-in=1,000. Then we chose a value t ranged from 0 to 1 as our threshold for method BayesB, i.e. if one's posterior probability was larger than t then the corresponding marker was included in the large effect class.

In this simulation study, we considered a situation existing highly correlations among large effect markers. As the simulation part in genomic selection, the phenotypic values were generated using the model 3.2.1. Two scenarios were displayed in this part. In Scenario I, 21 fixed markers were assigned different constants, i.e. (β_5, β_{664}, β_{1098}, β_{1099}, ..., β_{1115}, β_{1116}) = (-2.6, 0.5, -5.8, -2.2, -1.8, 2, 9, -2.1, -8.4, -3.4, 5.6, -10, -1.2, 3.3, 1.5, 7, -2.4, -0.3, 6, 0.8, 4). Figure 3.7 depicts the genetic effect against marker position. It obviously illustrates negligible, moderate, or sizable values were randomly assigned to genetic marker

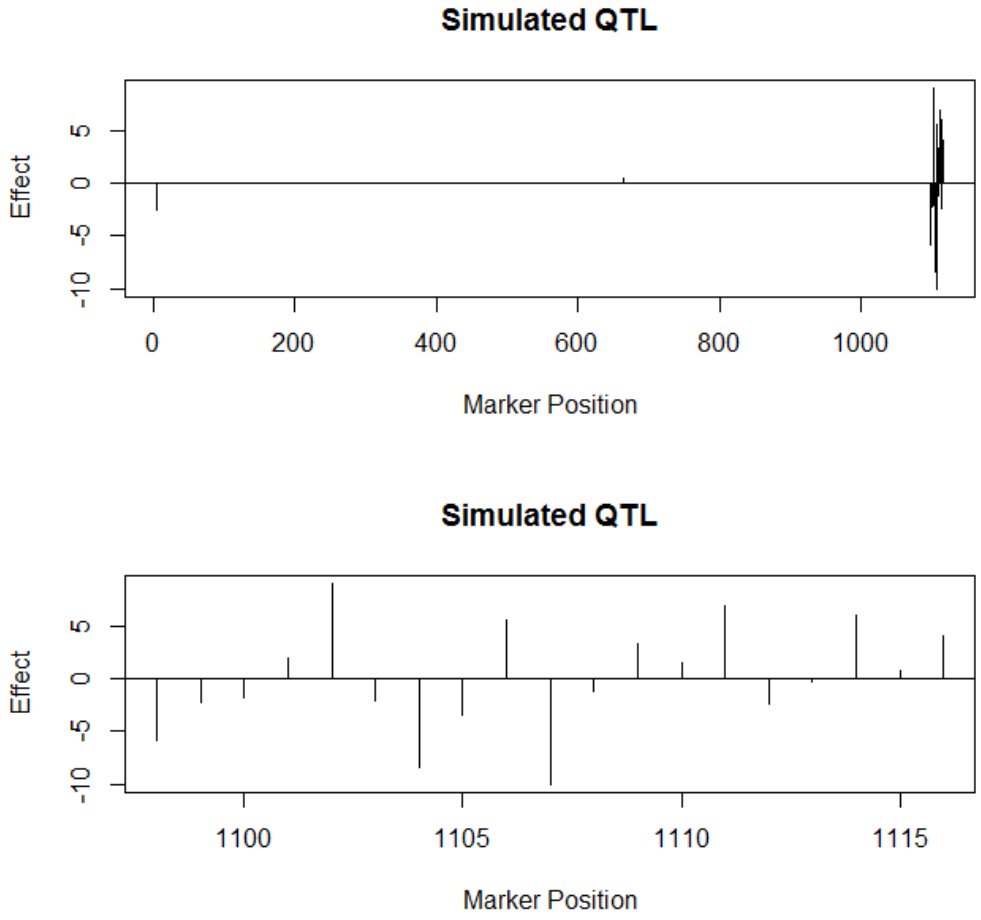

Figure 3.7 Simulated QTL effects and their marker positions. Upper plot exhibits all 21 markers and their genetic effects. Lower plot exhibits 19 highly correlated markers and their genetic effects.

effects. Note that except marker 5 and 664, markers through position 1098 to position 1116 were collinearity. The remaining marker effects were all identical to zero. The error term was sampled from $\mathrm{N}(0, I\sigma^2)$, where $\sigma^2 = (1, 3, 5, 7, 11, 15, 21)$. To simulate a more realistic situation, we considered that the genetic variations consisted of a large number of small effects and a small number of large effects. Therefore, in Scenario II, small marker effects were sampled from $\mathrm{N}(0, I0.1)$, and large marker effects and the error term were identical to the values in Scenario I.

In order to use the proposed approach in GWAS, we need to do an additional step rather than the steps in genomic selection: calculate the total selected count for each marker which is how many times this marker will be selected into the large effect cluster through the entire iteration process. If we define the total selected count of the i^{th} locus as c_i, then it can be calculated by

$$c_i = l - \sum_{j=1}^{l} u_{ij},$$

where l is the total number of iterations, and u_{ij} is the cluster label for the i^{th} marker in the j^{th} iteration. If one marker's total selected count is greater than threshold$\times l$, then we treat it as a large effect marker.

False discovery rate (FDR) and false negative rate (FNR) were employed to measure the model performance. They are defined as follows

$$\mathrm{FDR} = \frac{\text{the number of markers with } \hat{\beta}_i \neq 0 \text{ but } \beta_i = 0}{\text{the number of markers with } \hat{\beta}_i \neq 0},$$

and

$$\mathrm{FNR} = \frac{\text{the number of markers with } \hat{\beta}_i = 0 \text{ but } \beta_i \neq 0}{\text{the number of markers with } \beta_i \neq 0},$$

where $\hat{\beta}_i$ is the i^{th} predicted marker effect, β_i is the i^{th} actual marker effect, and $i = 1, \ldots, m$. Note that we didn't use false positive rate (FPR) because of large misclassification rates. Namely, the numerator value of FPR is greater than its denominator value, then FPR is greater than 1. FPR can be obtained by

$$\text{FDR} = \frac{\text{the number of markers with } \hat{\beta}_i \neq 0 \text{ but } \beta_i = 0}{\text{the number of markers with } \beta_i = 0}.$$

Figure 3.8 and figure 3.9 illustrate FDR and FNR against the σ^2 values in two scenarios. According to their definitions, FDR measures how likely a model overselects insignificant markers, and FNR measures how likely a model doesn't select the correct markers. Both figures indicate that SREML is superior to LASSO and BayesB. FDR of LASSO is extremely high which implies LASSO makes variable selection more prone to large errors. And low FNR for LASSO implies it only selects few markers from the collinearity group. It seems that BayesB tends to be more sensitive to noises because its FNR in Scenario II is worse than the one in Scenario I. In Scenario II, FDR of BayesB straightly goes down as σ^2 increases. The reason is it tends to select less markers if the threshold determining large effect markers is consistent. Table 3.5, table 3.6, table 3.7, and table 3.8 list the standard deviations of FDR and FNR under three models. Based on them, LASSO is the most robust one among three models. In Scenario II, FDR and FNR for SREML are more spread out than LASSO and BayesB.

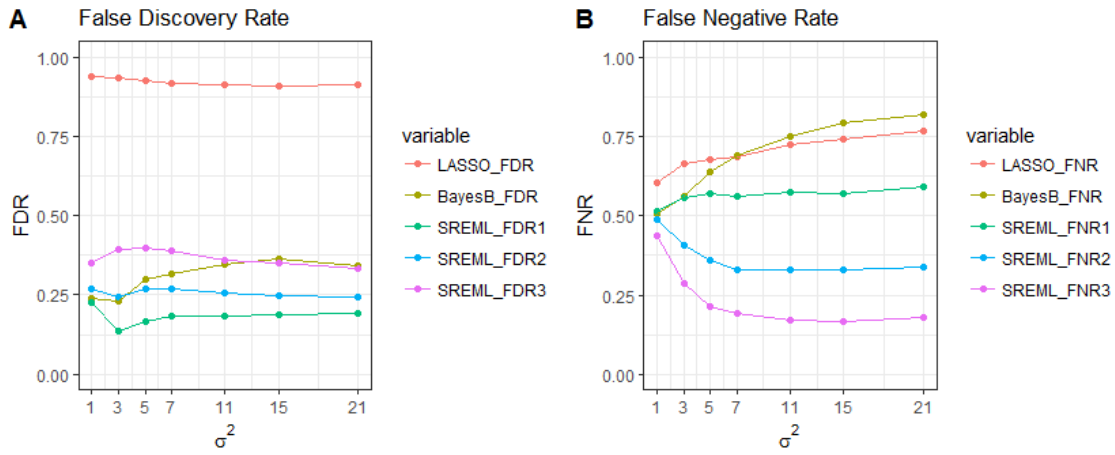

Figure 3.8 Comparison of FDR and FNR among SREML, LASSO, and BayesB in Scenario I. SREML_XXX1, SREML_XXX2, SREML_XXX3 represent the thresholds used for SREML in this simulation study were 0.5, 0.45, and 0.4, respectively. The threshold for BayesB was 0.2. The number of replicates per σ^2 was 250.

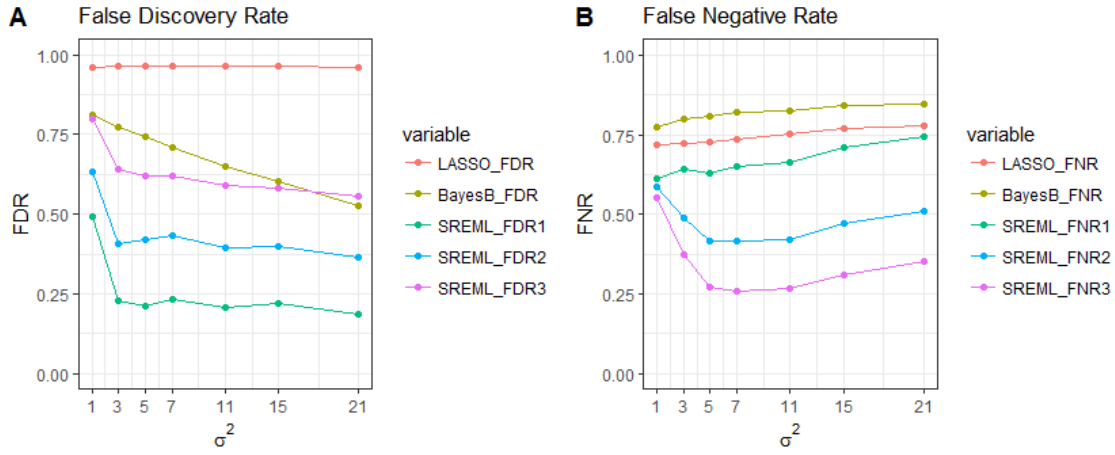

Figure 3.9 Comparison of FDR and FNR among SREML, LASSO, and BayesB in Scenario II. SREML_XXX1, SREML_XXX2, SREML_XXX3 represent the thresholds used for SREML in this simulation study were 0.5, 0.45, and 0.4, respectively. The threshold for BayesB was 0.3. The number of replicates per σ^2 was 250.

Table 3.5 Comparison of the standard deviation for FDR under three models in Scenario I. SREML1, SREML2, SREML3 represent the thresholds used for SREML were 0.5, 0.45, and 0.4, respectively. The threshold for BayesB was 0.2. The number of replicates per σ^2 was 250.

σ^2	LASSO	BayesB	SREML1	SREML2	SREML3
1	0.0081	0.1389	0.0872	0.0797	0.0708
3	0.0123	0.1307	0.0904	0.0831	0.0908
5	0.0165	0.1256	0.0945	0.0730	0.0839
7	0.0190	0.1296	0.0949	0.0708	0.0902
11	0.0217	0.1402	0.1074	0.0781	0.0892
15	0.0235	0.1529	0.0943	0.0720	0.0847
21	0.0255	0.1660	0.0923	0.0677	0.0741

Table 3.6 Comparison of the standard deviation for FDR under three models in Scenario II. SREML1, SREML2, SREML3 represent the thresholds used for SREML were 0.5, 0.45, and 0.4, respectively. The threshold for BayesB was 0.3. The number of replicates per σ^2 was 250.

σ^2	LASSO	BayesB	SREML1	SREML2	SREML3
1	0.0038	0.0439	0.1019	0.0627	0.0329
3	0.0048	0.0676	0.1333	0.0860	0.0525
5	0.0055	0.0780	0.1267	0.0874	0.0543
7	0.0062	0.0962	0.1368	0.0925	0.0538
11	0.0070	0.1212	0.1475	0.1005	0.0608
15	0.0077	0.1209	0.1484	0.1106	0.0650
21	0.0086	0.1632	0.1583	0.1273	0.0817

Table 3.7 Comparison of the standard deviation for FNR under three models in Scenario I. SREML1, SREML2, SREML3 represent the thresholds used for SREML were 0.5, 0.45, and 0.4, respectively. The threshold for BayesB was 0.2. The number of replicates per σ^2 was 250.

σ^2	LASSO	BayesB	SREML1	SREML2	SREML3
1	0.0461	0.0690	0.0392	0.0425	0.0481
3	0.0474	0.0670	0.0837	0.1104	0.1167
5	0.0435	0.0686	0.0945	0.1216	0.1208
7	0.0457	0.0646	0.1235	0.1399	0.1291
11	0.0503	0.0601	0.1268	0.1352	0.1240
15	0.0503	0.0655	0.1242	0.1393	0.1157
21	0.0479	0.0567	0.1201	0.1370	0.1104

Table 3.8 Comparison of the standard deviation for FNR under three models in Scenario II. SREML1, SREML2, SREML3 represent the thresholds used for SREML were 0.5, 0.45, and 0.4, respectively. The threshold for BayesB was 0.3. The number of replicates per σ^2 was 250.

σ^2	LASSO	BayesB	SREML1	SREML2	SREML3
1	0.0270	0.0443	0.0451	0.0419	0.0427
3	0.0357	0.0525	0.0838	0.0724	0.0918
5	0.0410	0.0569	0.1053	0.0996	0.1002
7	0.0450	0.0541	0.1368	0.0925	0.0936
11	0.0492	0.0556	0.1106	0.1028	0.0968
15	0.0486	0.0483	0.1017	0.1073	0.0917
21	0.0508	0.0406	0.0993	0.1137	0.1026

Figure 3.10 and figure 3.11 show the probabilities that a marker was selected among replicates under three models for both scenarios. A lot of unimportant markers were selected by LASSO, but SREML tended to only select markers with true large effects.

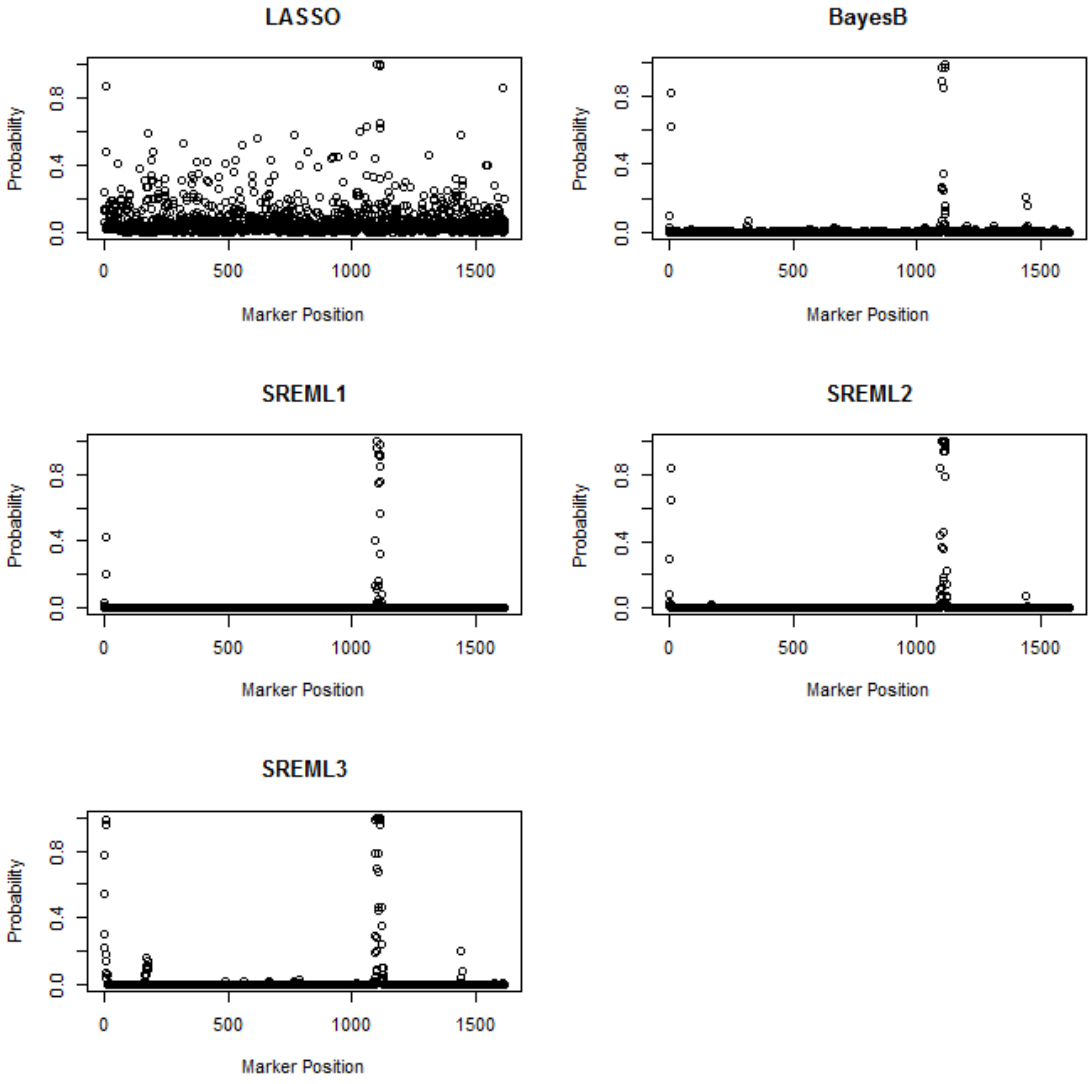

Figure 3.10 Comparison of probabilities among SREML, LASSO, and BayesB in Scenario I. SREML1, SREML2, SREML3 represent the thresholds used for SREML in this simulation study were 0.5, 0.45, and 0.4, respectively. The threshold for BayesB was 0.2. The number of replicates per σ^2 was 250.

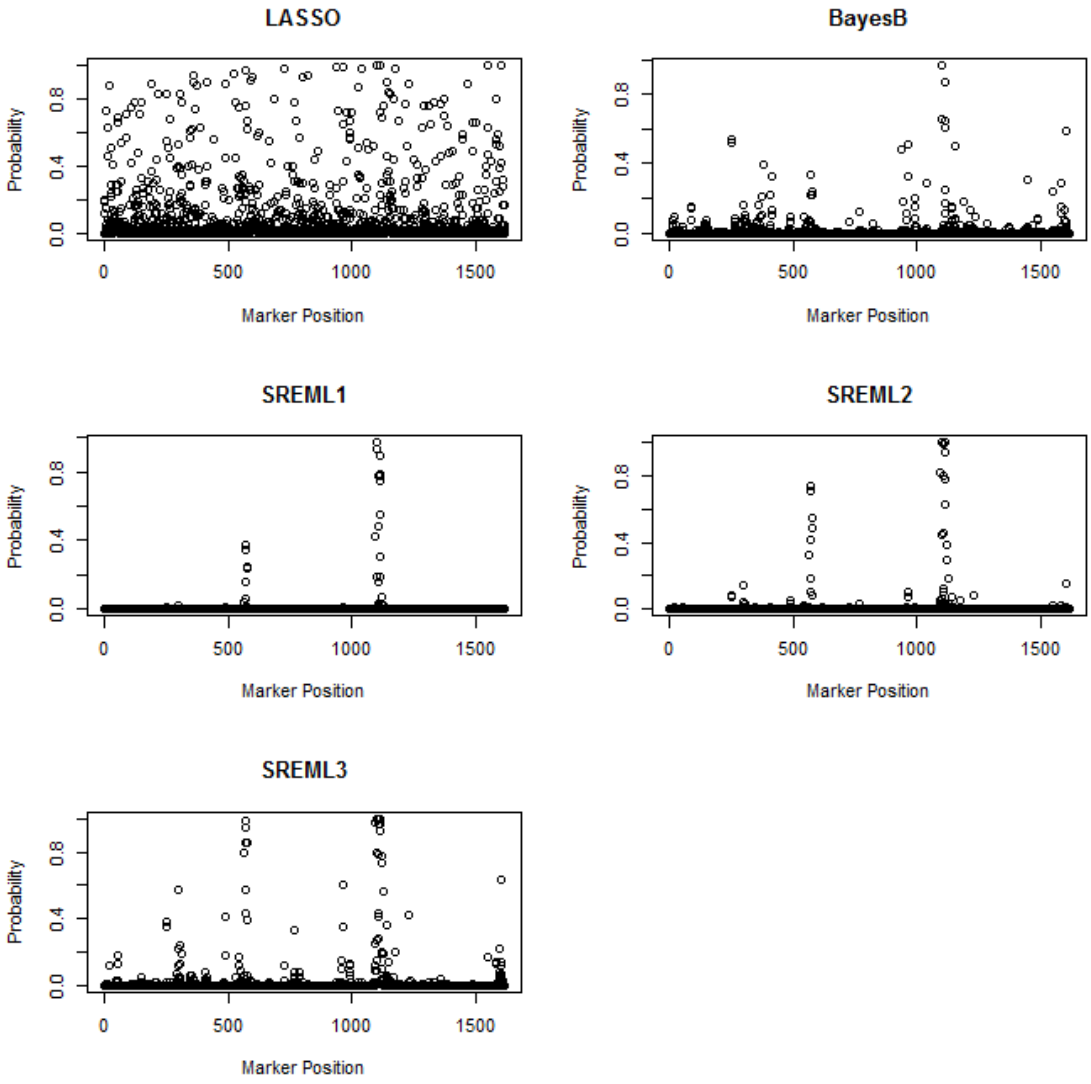

Figure 3.11 Comparison of probabilities among SREML, LASSO, and BayesB in Scenario II. SREML1, SREML2, SREML3 represent the thresholds used for SREML in this simulation study were 0.5, 0.45, and 0.4, respectively. The threshold for BayesB was 0.3. The number of replicates per σ^2 was 250.

Figure 3.7 illustrates marker effects change from marker to marker. Intuitively, markers with large effects are easier to be identified than markers with small effects. In fact, it is more important to identify markers with small effects. Here we consider effects between -3 and 3 as relatively small effects. Among 21 marker positions I listed above, 11 out of them can be treated as the relatively small effect group, which is S={5, 664, 1099, 1100, 1101, 1103, 1108, 1110, 1112, 1113, 1115}. Let us define true negative rate (TNR) within the relatively small effect group as

$$\text{TNR} = \frac{\text{the number of markers with } \hat{\beta}_i \neq 0 \text{ and } \beta_i \neq 0}{\text{the number of markers with } \beta_i \neq 0},$$

where $i \in S$. Table 3.9 lists TNR and its standard deviation under three models in Scenario II. SREML outperforms LASSO and BayesB, and its TNR gradually increases and then goes down as σ^2 increases. The performance of BayesB is poor since the underlying genetic architecture mismatches its model assumption.

Table 3.9 Comparison of TNR and its standard deviation under three models in Scenario II. SREML1, SREML2, SREML3 represent the thresholds used for SREML were 0.5, 0.45, and 0.4, respectively. The threshold for BayesB was 0.3. The number of replicates per σ^2 was 250.

σ^2	LASSO	BayesB	SREML1	SREML2	SREML3
1	0.0985(0.04)	0.0458(0.06)	0.1935(0.05)	0.2120(0.05)	0.2491(0.05)
3	0.0985(0.06)	0.0433(0.06)	0.1869(0.09)	0.3244(0.10)	0.4589(0.13)
5	0.0960(0.07)	0.0447(0.06)	0.2051(0.11)	0.4015(0.12)	0.5873(0.13)
7	0.0960(0.08)	0.0480(0.06)	0.2189(0.11)	0.4305(0.12)	0.6156(0.13)
11	0.0898(0.07)	0.0473(0.06)	0.1858(0.11)	0.4033(0.13)	0.5891(0.12)
15	0.0745(0.07)	0.0451(0.05)	0.1615(0.11)	0.3691(0.14)	0.5345(0.13)
21	0.0738(0.08)	0.0440(0.05)	0.1276(0.10)	0.3116(0.14)	0.4931(0.14)

3.4 Dicussion

A novel approach has been presented to enhance the accuracy of prediction in genomic selection and the accuracy of identifying QTL under collinearity in GWAS. BLUP assumes marker effects are randomly assigned from a normal distribution with homogeneous variance, while Bayesian alphabet models assume marker effect variances are heterogeneous. The proposed method SREML is between BLUP and Bayesian alphabet models. Unlike the assumption of BLUP, the novel method assumes a large group of markers have small effects and a small number of markers have extra effects than the remaining markers. This assumption allows we obtain additional information from the data to improve the model performance and obtain the markers with large genetic effects simultaneously. The simulation study in genomic selection shows the proposed method outperforms other methods when we used model 3.2.1 to generate phenotypic values.

As discussed above, the procedure of SREML is stochastic and it will be impressed that this algorithm is similar to Bayesian approaches. However, SREML is distinguishing from Bayesian approaches because we don't expect it will converge to a specific value. The result from each iteration is stochastic even through it is based on the result from the previous iteration. The reason is every element of the u vector is randomly sampled from a Bernoulli distribution with the posterior probability of success ρ, which involves a random variable π. And the value of π is randomly sampled from a Beta distribution and its parameters are calculated by taking advantage of the data information. It implies that the number of nonzero-effect markers is automatically obtained based on the data itself, but not an artificial setting. The unknown parameters are then estimated by the REML

function. Compared with Bayesian approaches, the cluster label vector u of SREML is the only random variable sampled from a distribution. Variance components are obtained according to the current cluster labels. Vector u will never converge. Therefore, the values of variance components vary from iteration to iteration. In addition, all unknown parameters are sampled from distributions in Bayesian approaches. Sometimes, the rate of convergence can be slow, which requires large iteration numbers. However, combining the stochastic step and the REML step in SREML facilitates to obtain good parameter estimates within the comparatively small iteration numbers.

The other contribution of this study is to make the algorithm to be computationally tractable. For each iteration, all markers need to be scanned to update the indicator variable u_i. It leads to the computational burden because the inverse and determinant of updated phenotypic variance V_i need to be calculated for every marker. Due to it, Woodbury matrix identities and the matrix determinant lemma are applied in the algorithm to reduce the computational efforts.

One application of GWAS is to identify QTL which are markers with large genetic effects. Nowadays, millions of SNPs can be produced as our genetic information, but only few of them are associated with quantitative traits. Therefore, a large amount of information is useless and time-consuming for GWAS. Selecting informative markers can achieve accuracy enhancement and marker interpretability, where SREML, LASSO, and BayesB make contributions on excluding redundant markers. Based on the results above, SREML outperforms LASSO and BayesB under collinearity condition in GWAS. LASSO tends to select unimportant markers than the other two methods, and LASSO and BayesB tends to select fewer of correlated markers than SREML. Since the genotypic values are highly

correlated, the posterior probabilities of u_i and u_j tends to be close to each other if their current labels are the same. Consequently, for this case, SREML is superior to LASSO and BayesB.

3.A Appendix of Chapter 3

A The R code for SREML

```
SREML<-function(z,y,w,iternum){

  n<-ncol(z)
  m<-nrow(z)
  x<-matrix(1,n,1)
  m1<-m*w
  m2<-m*(1-w)
  delta<-sample(c(rep(-1,m1),rep(1,m2)))
  pred<-NULL
  delt<-NULL
  max.pred<- -1e10
  max.delta<-delta
  max.iter<- 0
  max.PRESS<-0

  for(iter in 1:iternum){
```

```
z1<-z[which(delta==-1),]

z2<-z[which(delta== 1),]

k1<-crossprod(z1)

k2<-crossprod(z2)

kk<-list(k1,k2)

p<-length(kk)

fit<-mixed(x,y=y,kk) #REML method to estimate unknown parameters

fn0<- -fit[[1]]$fn

v1<-fit[[2]]$v1

v2<-fit[[2]]$v2

ve<-fit[[2]]$ve

parm<-c(v1,v2,ve)

v<-diag(n)*parm[p+1]

g<-matrix(0,n,n)

vv<-0

for(k in 1:p){

   vv<-vv+parm[k]

   g<-g+kk[[k]]*vv

}

v<-v+g

vi<-solve(v)

H<-g%*%vi
```

```
blup0<-y-x%*%fit[[1]]$beta
blup<-H%*%blup0
PRESS<-sum(((blup0-blup)/(1-diag(H)))^2)
SS<-sum((blup0-mean(blup0))^2)
PRED<-1-PRESS/SS
if(PRED>max.pred){
  max.pred<-PRED
  max.delta<-delta
  max.iter<-iter
  max.PRESS<-PRESS
  max.parm<-parm
  max.beta<-fit[[1]]$beta
}
b<-blup0
bb<-t(b)%*%vi%*%b
xx<-t(x)%*%vi%*%x
dv<-unlist(determinant(v))
dv<-dv[1]
alpha<-length(which(delta==-1))
beta<-m-alpha
w<-rbeta(1,alpha+m,beta+0.1*m)
pred<-rbind(pred,c(iter,alpha,w,fn0,PRESS,PRED,v1,v1+v2,ve))
```

```
      delt<-rbind(delt,c(iter,delta))
      for(i in 1:m){
        zk<-as.matrix(z[i,])
        zz<-t(zk)%*%vi%*%zk
        bz<-t(b)%*%vi%*%zk
        zb<-t(bz)
        xz<-t(x)%*%vi%*%zk
        zx<-t(xz)
        xHx<-xx-xz%*%solve(zz-1/(delta[i]*v2))%*%zx
        bHb<-bb-bz%*%solve(zz-1/(delta[i]*v2))%*%zb
        zvz<-abs(-delta[i]*v2*zz+1)
        dx<-unlist(determinant(xHx))
        dx<-dx[1]
        dv2<-log(zvz)
        fn<- -0.5*(dv+dv2+dx+bHb)
        u<-rbinom(1, 1, w/(w+(1-w)*exp(delta[i]*(fn0-fn))))
        delta[i]<-1-2*u
      }
    }
}

mixed<-function(x,y,kk){
```

```
loglike<-function(parm){
  v<-diag(n)*parm[p+1]
  vv<-0
  for(k in 1:p){
    vv<-vv+parm[k]
    v<-v+kk[[k]]*vv
  }
  vi<-solve(v)
  xx<-t(x)%*%vi%*%x
  xy<-t(x)%*%vi%*%y
  b<-solve(xx,xy)
  d1<-unlist(determinant(v))
  d1<-d1[1]
  d2<-unlist(determinant(xx))
  d2<-d2[1]
  r<-y-x%*%b
  q<-t(r)%*%vi%*%r
  loglike<- -0.5*(d1+d2+q)
  return(-loglike)
}
fixed<-function(parm){
```

```
  v<-diag(n)*parm[p+1]
  g<-matrix(0,n,n)
  vv<-0
  for(k in 1:p){
    vv<-vv+parm[k]
    g<-g+kk[[k]]*vv
  }
  v<-v+g
  vi<-solve(v)
  xx<-t(x)%*%vi%*%x
  xy<-t(x)%*%vi%*%y
  covb<-solve(xx)
  beta<-solve(xx,xy)
  yhat<-g%*%vi%*%(y-x%*%beta)
  yobs<-y-x%*%beta
  r2<-cor(yobs,yhat)^2
  result<-list(beta,covb,r2)
  return(result)
}
loglike0<-function(x,y){
  xx<-t(x)%*%x
  xy<-t(x)%*%y
```

```
  b<-solve(xx,xy)
  r<-y-x%*%b
  s2<-drop(t(r)%*%(r))/(n-ncol(x))
  v<-diag(n)*s2
  vi<-diag(n)/s2
  d1<-unlist(determinant(v))
  d1<-d1[1]
  d2<-unlist(determinant(xx))
  d2<-d2[1]
  q<-t(r)%*%vi%*%r
  loglike<- -0.5*(d1+d2+q)
  return(-loglike)
}
n<-length(y)
p<-length(kk)
fn0<-loglike0(x,y)
parm0<-rep(1,p+1)
result<-optim(par=parm0,fn=loglike,hessian = TRUE,
method="L-BFGS-B",lower=1e-5,upper=1e5)
parm<-result$par
conv<-result$convergence
fn<-result$value
```

```
  lrt<-2*(fn0-fn)
  hess<-result$hessian
  covp<- solve(hess)
  bb<-fixed(parm)
  beta<-bb[[1]]
  covb<-bb[[2]]
  r2<-bb[[3]]
  fixed<-data.frame(conv,fn0,fn,lrt,beta,covb,r2)
  v1<-parm[1]
  v2<-parm[2]
  ve<-parm[3]
  parm<-data.frame(v1,v2,ve)
  result<-list(fixed,parm)
  return(result)
}
```

B Supplemental tables

Table 3.10 Variance components estimated by SREML and BLUP in Scenario II.

Method	σ^2_γ	σ^2_S	σ^2_L	σ^2_e
SREML	-	8.8391	104.9971	1.4338
BLUP	22.2774	-	-	1.4377

Bibliography

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3):469–475.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127.

Bernardo, R. and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, 47(3):1082–1090.

Celeux, G. (1985). The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational statistics quarterly*, 2:73–82.

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., Li, Y., Liu, X., Zhang, H., Dong, H., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature genetics*, 46(7):714.

Cheng, H., Qu, L., Garrick, D. J., and Fernando, R. L. (2015). A fast and efficient gibbs sampler for bayesb in whole-genome analyses. *Genetics Selection Evolution*, 47(1):80.

Consortium, C. A. D. C. G. et al. (2011). A genome-wide association study in europeans and south asians identifies five new loci for coronary artery disease. *Nature genetics*, 43(4):339.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365):169–174.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., Camacho-González, J. M., Pérez-Elizalde, S., Beyene, Y., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in plant science*.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2):327–345.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Desta, Z. A. and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in plant science*, 19(9):592–601.

Domarkienė, I., Pranculis, A., Germanas, Š., Jakaitienė, A., Vitkus, D., Dženkevičiūtė, V., Kučinskienė, Z., and Kučinskas, V. (2013). Rtn4 and fbxl17 genes are associated with coronary heart disease in genome-wide association analysis of lithuanian families. *Balkan Journal of Medical Genetics*, 16(2):17–22.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., Bowman, P., Reich, C., Mason, B., and Goddard, M. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of dairy science*, 95(7):4114–4129.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Ghassibe-Sabbagh, M., Haber, M., Salloum, A. K., Al-Sarraj, Y., Akle, Y., Hirbli, K., Romanos, J., Mouzaya, F., Gauguier, D., Platt, D. E., et al. (2014). T2dm gwas in the lebanese population confirms the role of tcf7l2 and cdkal1 in disease susceptibility. *Scientific reports*, 4:7351.

Gianola, D. and Schön, C.-C. (2016). Cross-validation without doing cross-validation in genome-enabled prediction. *G3: Genes, Genomes, Genetics*, 6(10):3107–3128.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Harris, B., Johnson, D., Spelman, R., et al. (2009). Genomic selection in new zealand and the implications for national genetic evaluation. *ICAR Technical Series*, (13):325–330.

Hayes, B. J., Bowman, P. J., Chamberlain, A., and Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science*, 92(2):433–443.

Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Science*, 49(1):1–12.

Henderson, C. R. (1950). Estimation of genetic parameters. In *Biometrics*, volume 6, pages 186–187. INTERNATIONAL BIOMETRIC SOC 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210.

Henderson, C. R. (1963). Selection index and expected genetic advance. *Statistical genetics and plant breeding*, 982:141–163.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447.

Henderson, C. R., Kempthorne, O., Searle, S. R., and Von Krosigk, C. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15(2):192–218.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Huang, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., Li, M., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature genetics*, 42(11):961.

Huang, X., Yang, S., Gong, J., Zhao, Y., Feng, Q., Gong, H., Li, W., Zhan, Q., Cheng, B., Xia, J., et al. (2015). Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nature communications*, 6:6258.

Jia, G., Huang, X., Zhi, H., Zhao, Y., Zhao, Q., Li, W., Chai, Y., Yang, L., Liu, K., Lu, H., et al. (2013). A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (setaria italica). *Nature genetics*, 45(8):957.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

Lee, S. H., van der Werf, J. H., Hayes, B. J., Goddard, M. E., and Visscher, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome snp data. *PLoS genetics*, 4(10):e1000231.

Legarra, A., Robert-Granié, C., Manfredi, E., and Elsen, J.-M. (2008). Performance of genomic selection in mice. *Genetics*, 180(1):611–618.

Li, Y. R., Li, J., Zhao, S. D., Bradfield, J. P., Mentch, F. D., Maggadottir, S. M., Hou, C., Abrams, D. J., Chang, D., Gao, F., et al. (2015). Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nature medicine*, 21(9):1018.

Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92.

Pérez, P. and de Los Campos, G. (2014). Genome-wide regression & prediction with the bglr statistical package. *Genetics*, pages genetics–114.

Picard, R. R. and Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583.

Piepho, H.-P. (2009). Ridge regression and extensions for genomewide selection in maize. *Crop Science*, 49(4):1165–1176.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517.

Sim, X., Ong, R. T.-H., Suo, C., Tay, W.-T., Liu, J., Ng, D. P.-K., Boehnke, M., Chia, K.-S., Wong, T.-Y., Seielstad, M., et al. (2011). Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from southeast asia. *PLoS genetics*, 7(4):e1001363.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

VanRaden, P., Van Tassell, C., Wiggans, G., Sonstegard, T., Schnabel, R., Taylor, J., and Schenkel, F. (2009). Invited review: Reliability of genomic predictions for north american holstein bulls. *Journal of dairy science*, 92(1):16–24.

Verbyla, K. L., Hayes, B. J., Bowman, P. J., and Goddard, M. E. (2009). Accuracy of genomic selection using stochastic search variable selection in australian holstein friesian dairy cattle. *Genetics research*, 91(5):307–311.

Wei, G. C. and Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721.

Xu, S. (2017). Predicted residual error sum of squares of mixed models–an application to genomic prediction. *G3: Genes, Genomes, Genetics*, pages g3–116.

Xu, S., Xu, Y., Gong, L., and Zhang, Q. (2016). Metabolomic prediction of yield in hybrid rice. *The Plant Journal*, 88(2):219–227.

Xu, S., Zhu, D., and Zhang, Q. (2014). Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proceedings of the National Academy of Sciences*, 111(34):12456–12461.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565.

Yu, H., Xie, W., Wang, J., Xing, Y., Xu, C., Li, X., Xiao, J., and Zhang, Q. (2011). Gains in qtl detection using an ultra-high density snp map based on population sequencing relative to traditional rflp/ssr markers. *PloS one*, 6(3):e17595.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203.

Zhao, K., Tung, C.-W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., Norton, G. J., Islam, M. R., Reynolds, A., Mezey, J., et al. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in oryza sativa. *Nature communications*, 2:467.

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.